Text Mining and Real-Time Analytics of Twitter Data: A Case Study of Australian Hay Fever Prediction

Sudha Subramani[1(✉)], Sandra Michalska[1], Hua Wang[1], Frank Whittaker[1], and Benjamin Heyward[2]

[1] Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Australia
{sudha.subramani1,sandra.michalska}@live.vu.edu.au
[2] Nexus Online Pty Ltd., Greenock, Australia

Abstract. Social media platforms such as Twitter contain wealth of user-generated data and over time has become a virtual treasure trove of information for knowledge discovery with applications in healthcare, politics, social initiatives, to name a few. Despite the evident benefits of tweets exploration, there are numerous challenges associated with processing such data, given tweets specific characteristics. The study provides a brief of steps involved in manipulation Twitter data as well as offers the examples of the machine learning algorithms most commonly used in text analysis. It concludes with the case study on the Australian hay fever prediction with the application of the selected techniques described in the brief. It demonstrates an example of Twitter real-time analytics for heath condition surveillance with the use of interactive visualisations to assist knowledge discovery and findings dissemination. The results prove the potential of social media to play an important role in meaningful results extraction and guidance for decision makers.

Keywords: Twitter · Machine learning · Text mining
Information retrieval · Knowledge discovery

1 Introduction

Twitter is one of the most popular social media websites, where users can post and interact via posts called 'tweets' and it has been growing hastily since its creation in 2006 [1]. The platform's enormous benefit is the short time span that the messages can reach wide network of users, playing major role in real-time analytics [2]. Due to its ease of use, speed and reach, Twitter became a platform to set trends and agendas in topics that range from healthcare, through politics, technology, stock market analysis and entertainment industry. As Twitter has become a source for collective wisdom, many research studies used this power to predicting real-world outcomes. Twitter is also a cost effective and less time-consuming than other information extracting techniques such as surveys and opinion polls.

S. Siuly et al. (Eds.): HIS 2018, LNCS 11148, pp. 134–145, 2018.
https://doi.org/10.1007/978-3-030-01078-2_12

The enormous and high volume of information that disseminates through millions of Twitter users accounts presents an interesting opportunity to obtain a meaningful insight into population behavioural patterns along with the prediction of future trends. Moreover, gathering information on how people converse regarding topic can assist many sectors in the real-world applications.

In terms of case study selected, nearly 1 in 5 Australians suffered from allergic rhinitis in 2014 to 2015 [3]. The forecasts do not look promising due to climate changes as well as new allergens, worsening air quality etc. As the meteorological data on an array of hay fever triggers is becoming more and more available, there is currently no equivalent for the estimates of its prevalence and severity at the fine-grained spatial and temporal level. Thus, the study was inspired to fill this gap by utilising the real-time, low-cost and freely available social media to develop a proxy for pollen allergy prevalence and explore potential associations with the environmental factors.

The remainder of the paper is organised as follows. Section 2 presents the brief overview of text mining for Twitter application. Section 3 discusses various preprocessing techniques while dealing with noisy and unstructured data. Section 4 presents different classification based algorithms used in text mining. Section 5 introduces the real-time analytics of Twitter data and its applications in various domains are discussed. Section 6 ends with conclusions.

2 Text Mining for Twitter Application

The unstructured or semi-structured language is commonly used on Twitter or any other social media platform. Hence, the various types of ambiguities occur, such as morphological, syntactic or semantic. People tend to ignore grammatical rules and spelling mistakes in their posts [4]. In recent years, social media has become an active research area that has drawn huge attention among the research community for information retrieval and abstract topics discovery. Nonetheless, the following characteristics of Twitter makes it challenging for that purpose:

1. Immense volume, fast arriving rate and short message restriction,
2. Large number of spelling and grammatical errors,
3. Use of informal and mixed language,
4. High content of irrelevant data.

Therefore, an extraction of meaningful information from such noisy data became complex problem to solve. Text mining intends to address the above-mentioned issues. Liu et al. [5] defined text mining as an extension of data mining to text data. Information retrieval, text analysis, clustering and natural language processing are the multidisciplinary fields in text mining techniques. They facilitate models based on interesting patterns development and assist predictability.

3 Pre-processing Steps in Text Analysis

During data collection, the unstructured text data contains a lot of challenges that make it particularly challenging to work with as described in previous section. At the same time, the pre-processing steps are essential in any subsequent analyses. Precisely, if the data is not cleaned properly, the text analysis techniques at the later stage simply leads to "garbage in garbage out" phenomena [6]. Even though the pre-processing consumes a great amount of time, it improves the final output accuracy [7]. Feature extraction and feature selection are two basic methods of text pre-processing.

The content of collected tweets varies from useful and meaningful information to incomprehensible text. The former contains people's opinion and relevant posts regarding the topic, whereas the latter may contain advertisements and it does not add value to the analysis. Hence, high quality information and features are extracted by incorporating some pre-processing techniques explained briefly in the following subsections.

3.1 Feature Extraction

The Feature Extraction can be further categorized as 3 methods such as Morphological analysis, Syntactical analysis and Semantic analysis. The 3 categories are briefly explained below. The feature extraction is used for many applications such as automatic tweets classification [8], opinion analyser [9] and sentiment classification [10].

Morphological Analysis. Morphological analysis deals mainly with tokenization, stop-words removal and stemming [7]. The tokenization is the process of breaking a stream of text into words or phrases called tokens. Stop word lists contain common English words like articles, prepositions, pronouns, etc. Examples are 'a', 'an', the', 'at' etc. Hasan saif et al. [11] investigated that removing stop words improves the classification accuracy in Twitter analysis by reducing data sparsity and shrinking the feature space. Stemming is used to identify the root of a word, to remove the suffixes related to a term and to save a memory space. For example, the terms 'relations', 'related', 'relates' can be stemmed to simply 'relate'. Different stemming algorithms are available in the literature, such as brute-force, suffix-stripping, affix-removal, successor variety, and n-grams [7]. Porter stemming [12] is applied to standardise terms appearance and to reduce data sparseness. In addition to the above 3 methods, non-textual symbols and punctuation marks are removed. Noisy tweets are filtered by eliminating links, non-ascii characters, user mentions, numbers and hashtags.

Syntactical Analysis. Syntactic analysis consists of Part-of-Speech tagging (POS-tagging) and parsing techniques [13]. It provides knowledge about grammatical formation of the sentence and it is used to interpret logical meaning from the sentence. The POS tagging defines contextually related grammatical

sense in a sentence like noun, verb, adjective etc. Various approaches have been developed to implement POS tagging like Hidden Markov Model [13]. Parsing is another technique of syntactical analysis, where the sentence is represented in a tree-like structure and analysed for which group of words combine.

Semantic Analysis. Semantic analysis is the primary issue for relationship extraction form unstructured text [14]. This refers to wide range of processing techniques that identify and extract entities, facts, attributes, concepts and events to populate meta-data fields. This is usually based on two approaches like rule-based matching and machine learning approach. First approach is similar to entity extraction and requires the support of one or more vocabularies. Another one is machine learning approach and it deals with the statistical analysis of the content and derives relationship from the statistical co-occurrence of terms in the document corpus. WordNet-Affect [15] and SentiWordNet [16] are the popular approaches that are used to extract the useful contents from the textual message. Strapparava et al. [15] proposed the WordNet-Affect approach, a linguistic resource for a lexical representation of affective knowledge (affective computing is advancing as a field that allows a new form of human computer interaction in addition to the use of natural language). Another approach is SentiWordNet, which is proposed by Esuli et al. [16] and it is a publicly available lexical resource for opinion mining.

3.2 Feature Selection

Another essential step after feature extraction is feature selection that improves the scalability and accuracy of the classifier by constructing vector space. The main purpose of this approach is to select the most important subset of features from the original documents based on the highest score. The highest score is predetermined measure based on the importance of the word [17]. For the text mining, the high dimensionality of the feature space is the major hurdle, as it contains many irrelevant and noisy features. Hence Feature selection method is widely used to improve the accuracy and efficiency of the classifier. The selected features provide a good understanding of the data and retain original physical meaning.

A substantial amount of research has been applied to evaluate the predictability of features for the application in classification techniques. Among them, Peng et al. [18] studied how to select compact set of superior features at low cost according to a maximal statistical dependency criterion based on mutual information. Another approach is based on conditional mutual information and it is defined as a fast feature selection technique. This approach favours features that maximize their mutual information and ensures the selection of features that are both individually informative and 2-by-2 weakly dependent [19]. Mihalcea et al. [20] examined several measures to determine semantic similarity between short collections of text. It relies on simple lexical methods like pointwise mutual information and latent semantic analysis.

Another popular approach calculates feature vectors based on two basic methods, namely Term Frequency (TF) and Inverse Document Frequency (IDF). TF-IDF function is the combination of TF and IDF and is mainly used to estimate the frequency and relevancy of a given word in the document at the same time. Ramos et al. examined the results of applying TF-IDF to determine what terms in a corpus of documents might be more relevant to a query [21].

4 Literature Survey on Real-Time Analytics of Twitter Data

Twitter supports real time analytics in various aspects like spatial analytics, temporal analytics and text mining. Spatial analytics provides the visual representation of various trending topics across various geographical locations and temporal analysis presents an information about seasonal trends or outbreaks of various topics.

As for the examples, Kathy et al. [22] described a novel real-time flu and cancer surveillance system that uses spatial, temporal and text mining on Twitter data. The real-time analytics results are reported visually in terms of US disease surveillance maps, distribution and timelines of disease types, symptoms, and treatments. Several research studies focused on Twitter to analyse and predict sentiment analysis [23], opinion mining on political campaigns [24,25], natural disasters [26], epidemic surveillance [27], event detection [28], topic modeling [29–34], and so on. O'Connor et al. [25] and Tumasjan et al. [24] showed that sentiment analysis of tweets correlated with the voters' political preferences and closely aligned with the election results. Not only in the field of politics, but also in economics, have public tweets played a major role. Sentiment analysis has been previously studied on different aspects such as blogs and forums and has now been analysed in social media [35]. Bollen et al. [36,37] analysed that tweet sentiments can be used to predict trends of stock and it is directly correlated with them. Bruns et al. [38] and Gaffney et al. [39] observed that Twitter is a powerful tool to gather public opinion and create social change.

Sakaki et al. [26] investigated tweets during natural disasters and shown that it is able to detect earthquakes and send warning alerts to society. They considered each twitter user as a mobile sensor in Japan and the probability of an earthquake is computed using time and geolocation information of the user. Posting time and volume were modelled as exponential distribution to estimate locations of earthquake using kalman and particle filters. Their research further evidenced that earthquake can be sensed earlier than official broadcast.

Culotta et al. [40] analysed Twitter to detect influenza epidemic outbreaks that improves speed and cost reduction from traditional methods. Data of the user like gender, age and location can be used to provide more descriptive information about demographic insights compared to search queries. They detected influenza using multiple regression models and Quincey et al. [41] identified swine flu from Twitter using pre-defined keywords and terms co-occurrence method. These methods are analysed by searching the tweets with the keywords and

detected anomalous change with the rapid flow in message traffic related to given keywords. The aids of such a method is to collect more focused information from the Twitter stream. Twitter proves to be an effective source to research in healthcare topics and analyse various diseases like cholera [27], cardiac arrest [42], alcohol use [43], tobacco [44], drug use [45], mood swings [46] and Ebola outbreak [47]. Michael et al. proposed a technique called Ailment Topic Aspect Model [48,49] to monitor the health care of public through the diseases, symptoms and treatments detection in tweets.

Hence, this section describes the real-time application of Twitter in various sectors like healthcare, politics, natural disasters, stock market analysis, sentiment analysis and so on.

5 Case Study of Australian Hay Fever Prediction from Twitter

5.1 Experiment

The case study aiming to utilise machine learning algorithms to estimate the prevalence of Hay Fever from Twitter data was conducted. The steps involved relevant tweets extraction, followed by the standard pre-processing tasks, manual annotation, automatic classification with logistic regression model, correlation with the external data sources and statistical validation.

The tweets were extracted during high pollen season (mid-August up till end-November 2017) in Australia (location bounding box in the extraction criteria) and included either the 'hay fever' or 'hayfever' related terms or one of the associated with this condition symptoms (according to Wikipedia [50]).

The dataset of 681 tweets was manually annotated by the author, producing 402 Hay Fever - HF (59% of dataset) and Non-Hay Fever tweets - N-HF (41% of dataset). The logistic regression classifier was selected to train and test the data with the 3-times repeated 5-fold cross-validation. The TF-IDF frequency function was applied and the feature selection using filter method was adopted. The uni-grams were used based on the Minimum term frequency threshold set heuristically to 10.

Next, the potential predictors such as the weather condition variables, common triggers of pollen allergies, were identified and daily observations were collected. These in turn were correlated with HF tweeting intensity in set of locations (8 major Australian cities) on each day covering the analysed period. The Pearson's correlation coefficients for each city and weather variable on a daily temporal level were produced. For spatial patterns discovery and real-time analytics the interactive maps were developed.

5.2 Results

The first step after tweets extraction and pre-processing was training the classifier to automatically identify HF tweets from the collected dataset. The accuracy

obtained was 0.925 for 45 features based on the Minimum term frequency threshold of 10. Associated Kappa was 0.846.

Apart from high performance accuracy on a test dataset, the advantage of logistic regression classifier is an insight into the relevant terms used for prediction, thus allowing for any future selection criteria refinement. As the main goal is an overall system's sensitivity and precision maximisation, both extraction and classification form an integral part of a continuous improvement cycle.

The properly defined keywords allow to increase the ratio between the numbers of true positives to true negatives. Therefore, further investigation of terms identified as most predictive by the classifier and their corresponding coefficients enables better understanding of the classification criteria (Table 1). For instance, the word 'sneezing' was highly associated with HF related tweets, whereas 'allergy' occurred mostly in the false positives posts. Therefore, 'allergy' term is not recommended search query for future HF data extraction.

Table 1. Terms coefficients (shortlist).

Term	TF-IDF	Term	TF-IDF
Allergy	−2.47	Cat	−4.13
Reaction	0.82	Today	3.08
Sneezing	13.83	Itchy	0.74
Spring	2.74	Nose	1.83
Eye	0.23	Season	1.88

The words associations function further facilitated the knowledge discovery about the Hay Fever in Australia (Table 2). The combination 'watery eye' (r = 0.46) occurred with the correlation twice as high as either 'red eye' (r = 0.23) or 'swollen eye' (r = 0.22). The 'stuffy nose' (r = 0.50) was more common than 'itchy nose' (r = 0.22) and the 'sore throat' (r = 0.46) was the only meaningful and at the same time dominant association. The correlation score obtained is the degree of confidence in the word association. The values of the coefficients fall between 0.22 and 0.60 revealing moderate to strong correlation. The terms relevant to Hay Fever were underlined.

In terms of validation, the F test for the whole model proved statistically significant with $p < 0.001$. The highest adjusted r2 (adjusted for the number of predictors) was obtained for Melbourne (0.626). In other words, over 60% of the variance in the number of HF related tweets (as indicator of its prevalence) was able to be explained by the weather statistics and pollen data.

As pollen rates information for Melbourne covered only a proportion of the analysed period, the total number of observations included in the model was 67 (62.0% of the total). Regression constant was set to 0.

Finally, an interactive map to visually explore correlations of weather variables and HF tweeting intensity was developed. Strength and direction is

Table 2. Word associations and their corresponding correlation values for 'eye', 'nose' and 'throat' terms.

Eye		Nose		Throat	
Term	Correlation	Term	Correlation	Term	Correlation
watery	0.46	blow	0.6	sore	0.46
tired	0.31	bathroom	0.51	watery	0.32
itchy	0.3	stuffy	0.5	woke	0.3
red	0.23	hand	0.36	ear	0.26
drop	0.22	foundation	0.34	raw	0.26
barley	0.22	rub	0.34	thunderstorm	0.26
swollen	0.22	throat	0.25	heal	0.26
big	0.22	itchy	0.22	nose	0.25

Table 3. Multiple regression coefficients and p-values for Melbourne.

Statistic	Category	Unit	Completeness	Coef.	p-value
Min temperature	Temperature	Celsius	100.00%	−0.357**	0.045
Max temperature	Temperature	Celsius	100.00%	0.2	0.359
Ave temperature	Temperature	Celsius	100.00%	0.015	0.966
Rainfall	Precipitation	mm	100.00%	−0.156**	0.031
Evaporation	Precipitation	mm	100.00%	0.401***	0
Relative humidity	Precipitation	mm	100.00%	0.103***	0.007
Pressure	Pressure	hPa	100.00%	-0.007*	0.071
Max wind gust speed	Wind	km/h	100.00%	−0.037	0.334
Ave wind speed	Wind	km/h	100.00%	0.318***	0.001
Overcast	Overcast	oktas	100.00%	−0.296	0.162
Sunshine	Sunshine	hours	100.00%	−0.163	0.202
Pollen count	Pollen		62.00%	0.014	0.117

indicated by the size and colour gradient of the circle (orange indicates negative, whereas blue positive association) (Fig. 1).

5.3 Discussion

In terms of the analysis for Melbourne area, the moderately strong correlations were observed (Table 3). In particular, the positive correlation between the Average Wind Speed and HF tweeting pattern is worth noting as wind plays a major role in pollen grains spread, triggering the allergic symptoms. Another positive and significant correlation occurred for Evaporation and Relative Humidity. Usually, the plants are more likely to release their pollen into the air more on a sunny rather than rainy day. However, if the rain is occurring around a thunderstorm,

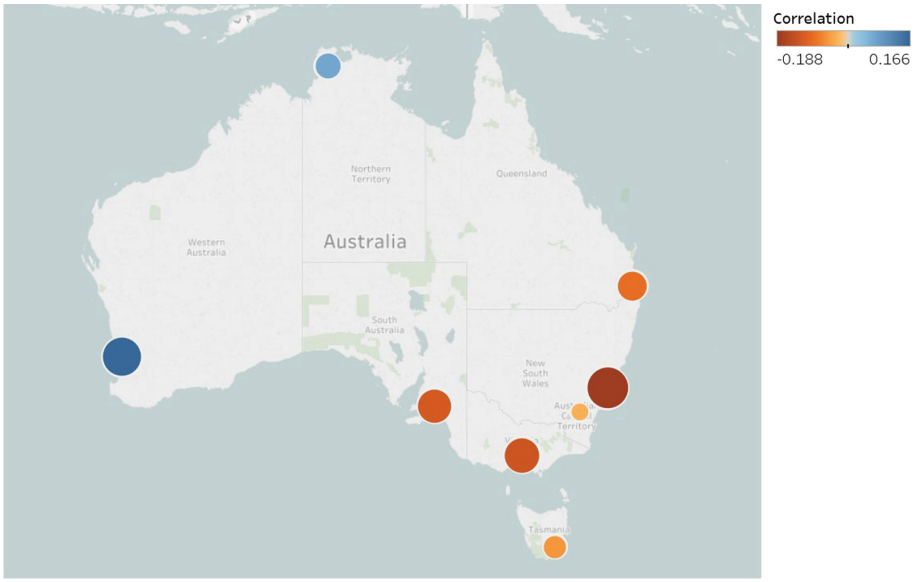

Fig. 1. Relative Humidity correlation map. (Color figure online)

then the humidity can make pollen grains burst open, releasing a high density of pollen into the air [51]. That may explain the coefficients values obtained. Furthermore, the Victoria State is known for its high probability of another co-related respiratory condition occurrence called thunderstorm asthma. As a matter of fact, the positive correlation for Relative Humidity variable paralleled the findings from the study on thunderstorm asthma predictability conducted in Melbourne reporting higher humidity with higher asthma admissions [52].

The correlation between pollen grains count and hay fever tweeting intensity was found insignificant ($p = 0.117$), although the value obtained was weakly positive (0.014). The pollen data for Melbourne was collected from 6 different pollen stations. In the analysis, the average was taken into account what might have affected the final output accuracy due to variations across the locations.

6 Conclusion

This survey provides the high-level overview of the specifics of Twitter data analysis, the challenges present as well as the current approaches to address them. The numerous applications utilising the real-time analytics potential from previous studies that transform unstructured tweets into valuable knowledge are given along with the case study on the Australian hay fever prediction. The experiment combined multiple heterogeneous data sources (numerical - structured vs text - unstructured) in order to obtain an instant insight into potential triggering factors of pollen allergy with the use of machine learning algorithms

as well as interactive maps. The correlation values obtained allowed to measure the impact of specific variables in order to assist future forecasting ability. Finally, an analysis of logistic regression outputs (terms coefficients magnitudes and directions) enabled further extraction and classification criteria refinement for an on-going real-time analysis in a continuous improvement cycle.

References

1. Twitter. https://about.twitter.com/company
2. Bruns, A., Stieglitz, S.: Towards more systematic twitter analysis: metrics for tweeting activities. Int. J. Soc. Res. Methodol. **16**(2), 91–108 (2013)
3. Australian Institute of Health and Welfare. Allergic Rhinitis ('Hay Fever') in Australia (2016)
4. Sorensen, L.: User managed trust in social networking-comparing Facebook, Myspace and Linkedin. In: 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, Wireless VITAE 2009, pp. 427–431. IEEE (2009)
5. Liu, F., Xiong, L.: Survey on text clustering algorithm-research present situation of text clustering algorithm. In: 2011 IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS), pp. 196–199. IEEE (2011)
6. Dai, Y., Kakkonen, T., Sutinen, E.: MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. **3**, 165–173 (2011)
7. Forman, G., Kirshenbaum, E.: Extremely fast text feature extraction for classification and indexing. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 1221–1230. ACM (2008)
8. Stavrianou, A., Brun, C., Silander, T., Roux, C.: NLP-based feature extraction for automated tweet classification. Interact. Data Min. Nat. Lang. Process. **145** (2014)
9. Zhao, P., Li, X., Wang, K.: Feature extraction from micro-blogs for comparison of products and services. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) WISE 2013. LNCS, vol. 8180, pp. 82–91. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41230-1_7
10. Shirbhate, A.G., Deshmukh, S.N.: Feature extraction for sentiment classification on twitter data. Int. J. Sci. Res. (IJSR), 2319–7064 (2016). ISSN (Online)
11. Saif, H., Fernández, M., He, Y., Alani, H.: On stopwords, filtering and data sparsity for sentiment analysis of twitter (2014)
12. Porter, M.F.: Snowball: a language for stemming algorithms (2001)
13. Yuan, L.: Improvement for the automatic part-of-speech tagging based on Hidden Markov Model. In: 2010 2nd International Conference on Signal Processing Systems (ICSPS), vol. 1, pp. V1–744. IEEE (2010)
14. Jadhao, H., Aghav, D.J., Vegiraju, A.: Semantic tool for analysing unstructured data. Int. J. Sci. Eng. Res. **3**(8) (2012)
15. Strapparava, C., Valitutti, A., et al.: WordNet affect: an affective extension of WordNet. In: LREC, vol. 4, pp. 1083–1086. Citeseer (2004)
16. Esuli, A., Sebastiani, F.: SentiWordNet: a high-coverage lexical resource for opinion mining. Evaluation **17**, 1–26 (2007)

17. Montañés, E., Fernández, J., Díaz, I., Combarro, E.F., Ranilla, J.: Measures of rule quality for feature selection in text categorization. In: R. Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 589–598. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45231-7_54
18. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
19. Fleuret, F.: Fast binary feature selection with conditional mutual information. J. Mach. Learn. Res. **5**(Nov), 1531–1555 (2004)
20. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI, vol. 6, pp. 775–780 (2006)
21. Ramos, J., et al.: Using TF-IDF to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning, vol. 242, pp. 133–142 (2003)
22. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1474–1477. ACM (2013)
23. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, pp. 36–44 (2010)
24. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. Icwsm **10**(1), 178–185 (2010)
25. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. Icwsm **11**(122–129), 1–2 (2010)
26. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
27. Chunara, R., Andrews, J.R., Brownstein, J.S.: Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian Cholera outbreak. Am. J. Trop. Med. Hyg. **86**(1), 39–45 (2012)
28. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 181–189 (2010)
29. Jiang, H., Zhou, R., Zhang, L., Wang, H., Zhang, Y.: A topic model based on Poisson decomposition. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1489–1498. ACM (2017)
30. Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., Zhang, X.: A probabilistic method for emerging topic tracking in microblog stream. World Wide Web **20**(2), 325–350 (2017)
31. Peng, M., Xie, Q., Wang, H., Zhang, Y., Tian, G.: Bayesian sparse topical coding. IEEE Trans. Knowl. Data Eng. (2018)
32. Peng, M., et al.: Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding. ACM Trans. Knowl. Discov. Data (TKDD) **12**(3), 38 (2018)
33. Peng, M., et al.: Neural sparse topical coding. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 2332–2340 (2018)

34. Yao, W., He, J., Wang, H., Zhang, Y., Cao, J.: Collaborative topic ranking: Leveraging item meta-data for sparsity reduction. In: AAAI, pp. 374–380 (2015)
35. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retr. **2**(1–2), 1–135 (2008)
36. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
37. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Icwsm **11**, 450–453 (2011)
38. Bruns, A., Burgess, J.E.: # Ausvotes: How twitter covered the 2010 Australian federal election. Commun. Polit. Cult. **44**(2), 37–56 (2011)
39. Gaffney, D.: iranElection: quantifying online activism. In: Proceedings of the Web Science Conference WebSci10. Citeseer (2010)
40. Culotta, A.: Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the First Workshop on Social Media Analytics, pp. 115–122. ACM (2010)
41. de Quincey, E., Kostkova, P.: Early warning and outbreak detection using social networking websites: the potential of twitter. In: Kostkova, P. (ed.) eHealth 2009. LNICST, vol. 27, pp. 21–24. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-11745-9_4
42. Bosley, J.C., et al.: Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation **84**(2), 206–212 (2013)
43. Culotta, A.: Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. Lang. Resour. Eval. **47**(1), 217–238 (2013)
44. Cobb, N.K., Graham, A.L., Byron, M.J., Niaura, R.S., Abrams, D.B., Participants, W.: Online social networks and smoking cessation: a scientific research agenda. J. Med. Internet Res. **13**(4) (2011)
45. Paul, M.J., Dredze, M.: Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 168–178 (2013)
46. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science **333**(6051), 1878–1881 (2011)
47. Odlum, M., Yoon, S.: What can we learn about the ebola outbreak from tweets? Am. J. Infect. Control. **43**(6), 563–571 (2015)
48. Paul, M.J., Dredze, M.: Discovering health topics in social media using topic models. PloS one **9**(8), e103408 (2014)
49. Paul, M.J., Dredze, M.: You are what you tweet: analyzing twitter for public health. Icwsm **20**, 265–272 (2011)
50. Allergic_rhinitis. https://en.wikipedia.org/wiki/Allergic_rhinitis
51. Allergy_cosmos. https://www.allergycosmos.co.uk/blog/why-is-my-hay-fever-worse-when-it-rains/
52. Silver, J.D., et al.: Seasonal asthma in Melbourne, Australia, and some observations on the occurrence of thunderstorm asthma and its predictability. PloS one **13**(4), e0194929 (2018)