Spatial and Temporal Word Spectrum of Social Media

Xue Li
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane 4072, Australia
xueli@itee.uq.edu.au

Guohun Zhu
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane 4072, Australia
g.zhu@uq.edu.au

Xiaolei Guo
China Academy of Electronics and Information Technology, Beijing, China
guoxiaolei247@hotmail.com

Weitong Chen
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane 4072, Australia
w.chen9@uq.edu.au

ABSTRACT

We assume that different users from different geolocations will post different microblog messages for their local issues on social networks. Based on this assumption, we calculate unique features of social media for different geolocations in different time periods. The outcome of this calculation is called a Spatial-Temporal Word Spectrum (STWS) model which is a *linguistic fingerprint* of a geolocation on social media. We use STWS as a baseline to catch the prominent and statistical features of microblogs as a spectral representation of the words used by social network users. As a baseline of the social media, STWS can be used to detect emerging local events. It can also be used to guess the location of a user if her/his location is unknown and the posted microblog exhibits the spectral features of that location. We show how STWS is visualized and how it is used to reveal behavioral features of local social network users. Our experiments show that the proposed method is effective and STWS opens a new way of studying social media.

Categories and Subject Descriptors

G.1.2 [**Mathematics of Computing**]: Approximation – *Special function approximations*

General Terms

Theory, Algorithms, Measurement, Performance.

Keywords

Special Temporal Word Spectrum, Social Media, Social Networks, Emerging Event Detection, Location Detection.

1. INTRODUCTION

It is common knowledge that different places have different lexical habits. In the theory of idiolect, individual persons have unique use of language to express ideas in their speeches [8]. People could detect the regions from the speaker's special words. These could also be presented in social media, such as Twitter, Facebook and Weibo. For example, a location indicative words method for predicting city names was presented in [3].

In addition, temporal and geographic information is strongly related to topic issues in social media [15]. For example, when an earthquake occurred, the topic on Twitter or Facebook would become a hot topic [9]. The classification of time zones of Twitter users was studied and applied in [14].

In social media analysis, such as opinion mining and sentiment analysis [4-6, 16], we need to know **when**, **where**, **what** and **who** are involved in the social media [19-21]. So a variety of applications can be support for people to make informed decisions.

This paper presents a novel method namely, Spatial-Temporal Word Spectrum (STWS) of social media to represent the unique temporal distributions of lexical habits in different regions. The regions are organized by postcodes or city names. The spectrum of words is statistical such that the words appeared in certain time and in specific areas. Like other information retrieval applications, this paper shows that the spatial and temporal word spectrum is applied in two different fields: emerging local event detections and geolocation predications.

2. RELATED WORK

Frequency spectrum is a concept in signal processing [17]. The values of the spectrum could be power or phase, which could be generated from *Fourier* transformation. The power spectrum density has been efficiently applied in signal processing [23]. However, it is based on time series and difficult to be applied in social media due to the lexical properties of text information.

The concept of word spectrum was initially used to visualize the word associations by Chris [12] using Google's Bi-Gram Data [13]. For example, "war memorial" occurs 531,205 times, while "peace memorial" occurs only 25,699 on their sample web pages [12]. The Google's word n-gram models are for a variety of research purposes, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others [1, 4-6, 16, 19-21]. In this paper we extend the concept of word spectrums with unique spatial-temporal information [2] for social media analysis and its applications.

SIGKDD Workshop WISDOM, August 10–13, 2015, Sydney, New South Wales, Australia.

3. PROPOSED METHODS

3.1 Spatial and Temporal Word Spectrum

Definition 1. *(STWS - Spatial and Temporal Word Spectrum)*

The Spatial and Temporal Word Spectrum (STWS) is a matrix $M = f(W, G, T)$*, where W is a set of words, G is a set of geographic locations, and T is an ordered set of time periods in hours (1-24). Function f(•) is defined by TF-IDF, where TF is the term frequency with respect of G and T for all given words, IDF (inverse document frequency) is also calculated based on G and T.*

In our approach, the concept of a document is regarded as a collection of words of a geolocation over a given period of time. Unlike the traditional TF-IDF concept, the row vectors of matrix *M* is restricted by both location as well as time period (see an example in Figure 1).

The calculation of *M* will need to be performed regularly and long enough on the entire social media to produce such a STWS for every geo-location, every frequently spoken word, in each time period (i.e., 1–24 hours). In our approach we consider time in hourly intervals.

In practice, only a very small proportion of microblogs posted by social network users would have geolocation tags available [19]. So the proposed STWS approach would need to collect data with not only a sufficient amount of data over a certain time period, but also a determination if the discovered STWS is stable (converging).

3.2 TF and IDF based on Regions

In general, term frequency (*TF*) is counted for a document *d*. In our context, *d* is set of microblogs collected from a particular geolocation within a time period (i.e., in an hour). *TF* is defined as *tf(t, h, p),* where *t* is a term, *h* is a time period of an hour, and *p* is a postcode.

The inverse document frequency (*IDF*) is defined as

$$idf(t) = \log(N / n(t)) \tag{1}$$

where *N* is the total postcodes in the datasets and *n*(*t*) is the number of postcodes where users posted term *t*. Thus, the *TF-IDF* is denoted as,

$$tfidf(t, h, p) = tf(t, h, p) * idf(t) \tag{2}$$

Thus together, terms, hours and postcodes build a STWS which represents a total statistical information about **where, when**, and **what** have been regularly happening on the social media.

3.3 Mutual Information for STWS

Mutual Information (*MI*) is a measure indictor of the variables' mutual dependence between time variable *X* and the geolocation (e.g., postcode) variable *Y*. The *MI*(*X*, *Y*) between pairwise parameters *X* and *Y* is the relative entropy between the join distribution and the product distribution *P*(*x*)*P*(*y*). It is denoted as,

$$MI(X, Y) = H(X) - H(X \mid Y) \tag{3}$$

where *H*(*X*) is the Shannon' entropy,

$$\begin{aligned} H(X) &= \sum_i P(x_i) I(x_i) \\ &= \sum_i P(x_i) \log_b P(x_i) \end{aligned} \tag{4}$$

where *H*(*X*|*Y*) is the conditional entropy,

$$H(X \mid Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \quad . \tag{5}$$

MI is used for measuring the correlations between spatial and temporal information in STWS.

4. Experimental results

4.1 Datasets

Two real world datasets are used to illustrate our proposed approach. The first one is named AU-TWEET, which we extracted from Twitter from May 2015 to June 2015. The language of Twitter was restricted to English. Only geo-tagged tweets for Australian geolocations were selected and the postcode information were obtained from *Foursquare*[1] and *Google Map*[2] because some geolocation couldn't found from *Foursquare*. The second one is extracted by Twitter Stream API for collecting the tweets from Sydney Metropolitan Area, from 25 June 2015 to 28 June 2015, and 37,225 Tweets in total.

4.2 Spatial Temporal Word Spectrum Graphs

We rank the high-frequency non-stop words as query topics based on locations and events from social media. Table 1 shows some randomly selected highly-ranked frequent words used in our experiment.

Table 1 Word set in experiment

Adelaide	*airport*	*beach*	*bed*	*breakfast*
Brisbane	*campus*	*coast*	*course*	*dinner*
dog	*god*	*gold*	*Griffith*	*gym*
hill	*hotel*	*hpa*	*humidity*	*lucia*
lunch	*mainland*	*Melbourne*	*night*	*nsw*
Perth	*Queensland*	*qut*	*rain*	*sleep*
storm	*Sydney*	*temperature*	*train*	*university*
Vic	*wind*			

Figure 1 shows a sample STWS model for some regions. We only used 14 postcodes for display. The x-axis is for different GMT hours of the day. It is shown that during the night (+10 Brisbane time), the *tf-idf* variations in *Toowoomba* area (postcode 4350) are small, which imply that most of residents in that area are in sleep due to the cold winter and the high mountains area. In contrast, people in Sydney are still busily tweeting because of the night living style.

Figure 2 shows the STWS variations between selected words and areas. The Melbourne city area contains the discussions on weathers, such as the terms of wind, temperature, etc. However there may be uncertainties to determine geolocations based on the unique features in an STWS, since weather can be discussed in anywhere. In fact, the weather topics in postcode 3000 (Melbourne city) area are always frequent and significantly higher

1 www.square.com

2 www.google.com/maps/

than other areas in most hours as shown in Figure 3. Conversely, the topic on dogs is frequent in afternoon between 15:00 – 16:00 o'clock. This phenomena indicates that for a given geolocation, we need to find a specific set of spatial-temporal and highly-ranked feature words to flavour a specific geolocation (i.e., the discovering of the *linguistic fingerprint*).

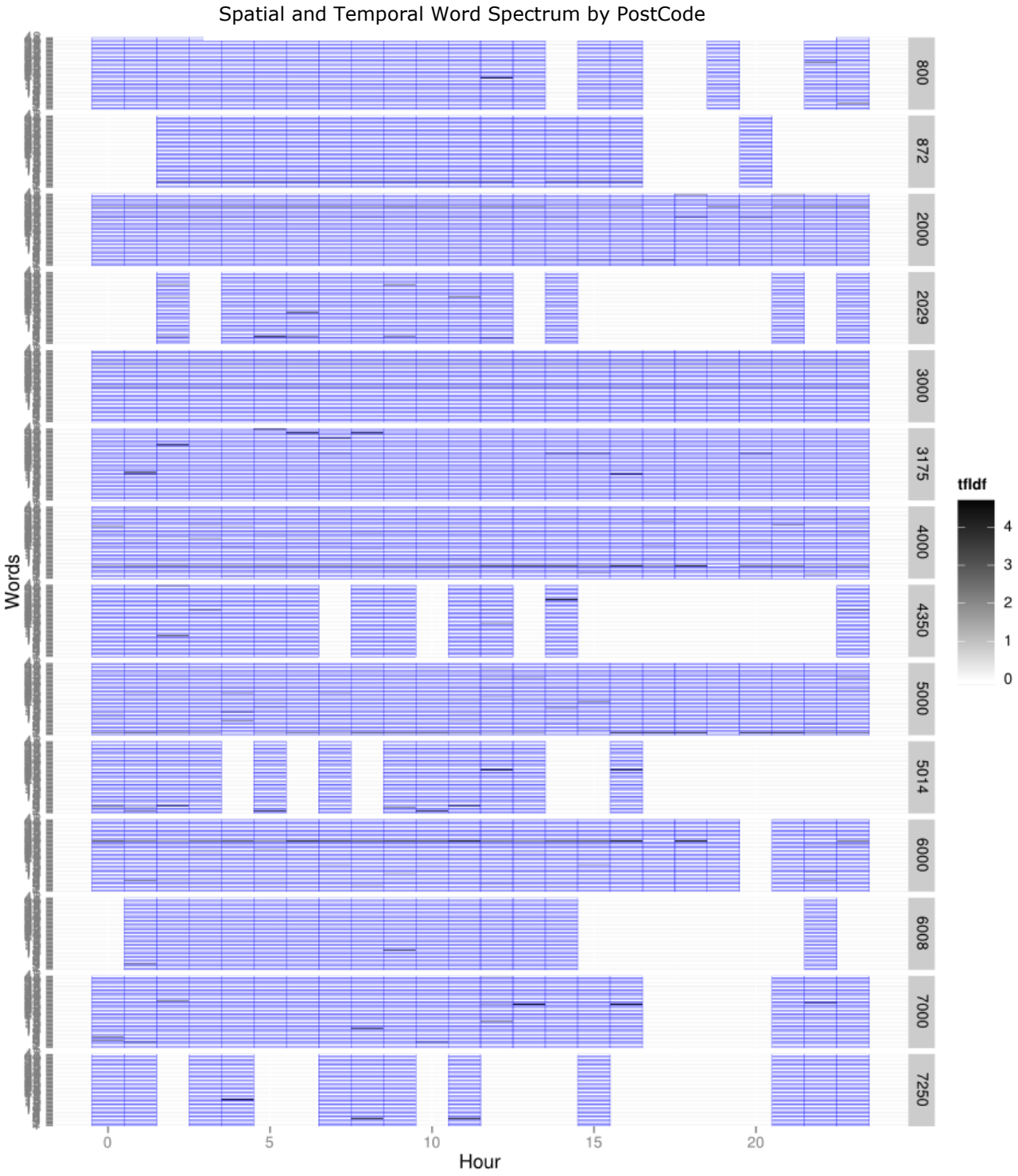

Figure 1. Spatial and Temporal Word Spectrum among words, hours, and regions

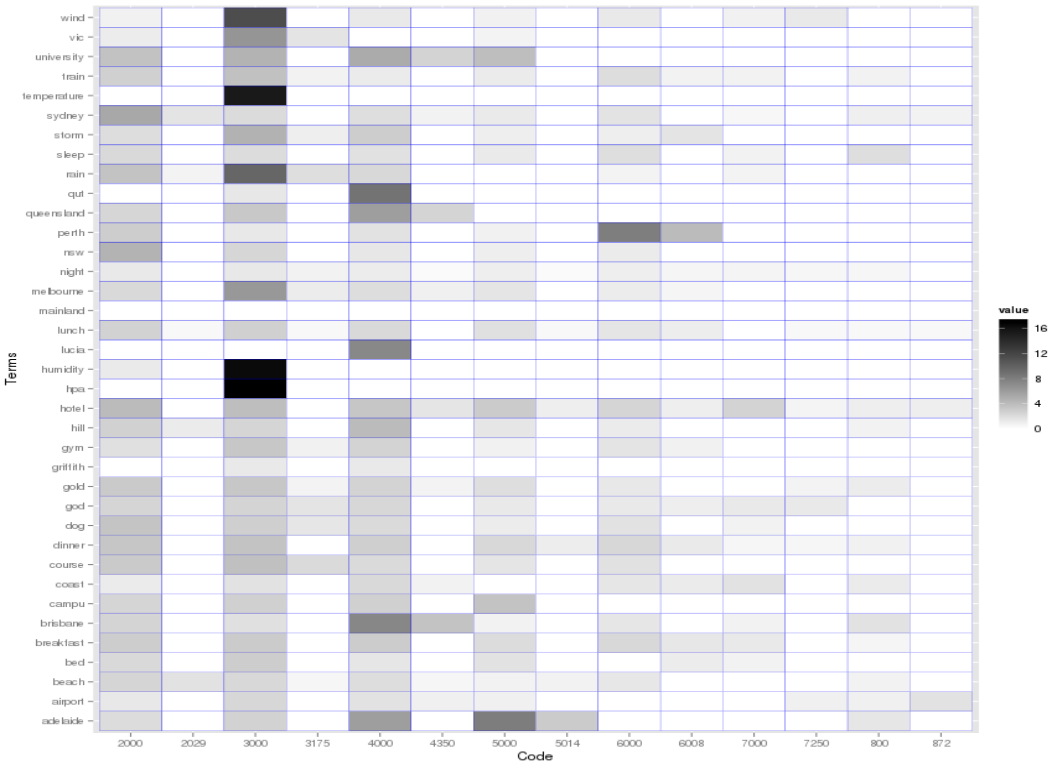

Figure 2. TF/IDF model of relationship between words and regions

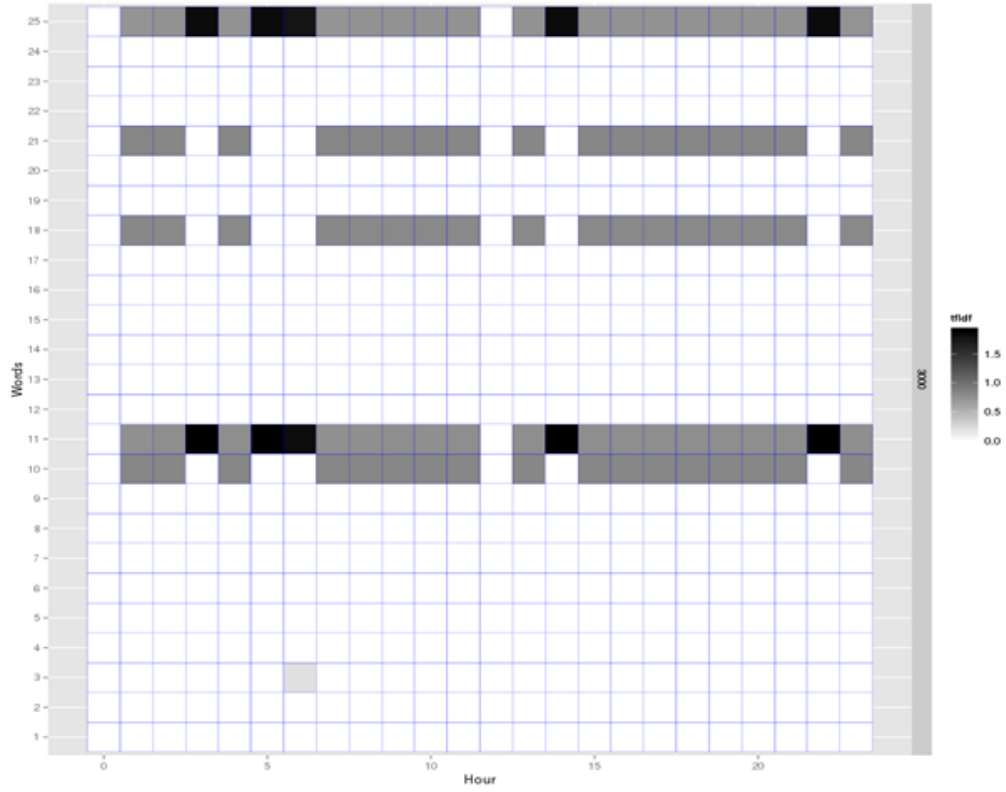

Figure 3. Spatial-Temporal Word Spectrum in postcode 3000 (Melbourne city) region

4.3 MI of Different Regions

Figure 4 shows the *MI* values of randomly selected 14 postcode with hourly 1-24 time periods. It is clearly seen that Brisbane area (postcode 4000) has the most differentiated topics among 24 hours. In contrast, one of the Melbourne areas (postcode 3175) has the lowest *MI*, which implies that those topics are difficult to distinguish among the time periods and between different geolocations.

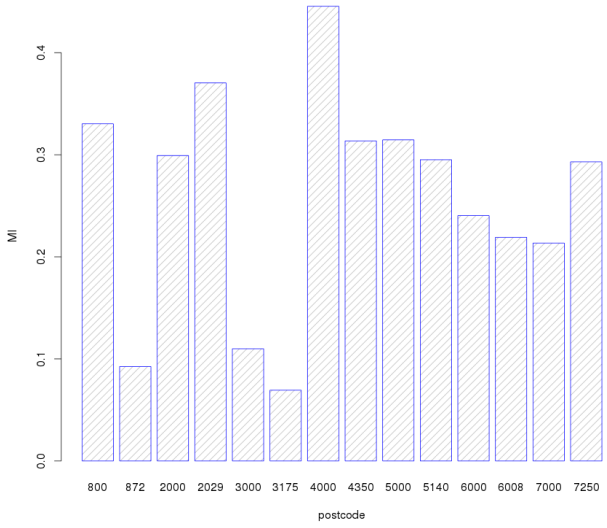

Figure 4. MI of different Postcodes

Figure 5 shows that the different numbers of tweets will be posted in different areas. This information is also useful to understand the property of a geolocation. The highest number of tweets is in Melbourne city area on 10:00 o'clock in morning, which is larger than the total number of the tweets posted within both Sydney and Brisbane areas. Then the numbers of tweets of those areas increase up to 21:00 o'clock in Brisbane, 19:00 o'clock in Melbourne. During the periods between 14:00–06:00 o'clock, the numbers of tweets of those areas are all at low levels.

Compared to Figure 4, it is shown that the diversity of Twitter contents and topics does not depend on the numbers of tweets. While in contrast, the numbers of tweets are strongly associated with the geolocations (i.e., related to the populations of those areas).

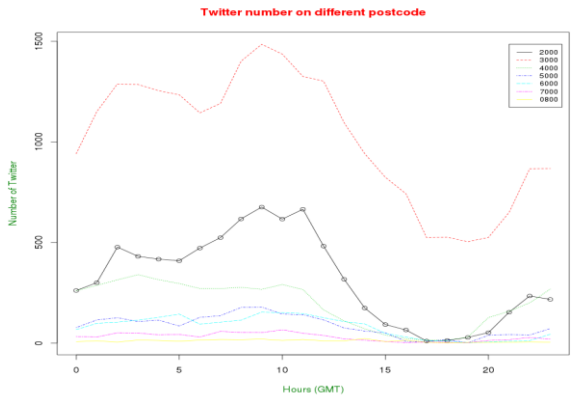

Figure 5. The numbers of tweets in different hours of a day.

5. APPLICATION OF STWS MODEL FOR EVENT DETECTION

For emerging local events detection on social media, one key problem is detecting the burst features. In many literatures [10, 22], the burst is detected by analysing the extent of the deviation between a word's current frequency and its usual frequencies. If there is a significant deviation, it is identified as a bursty word. Usually, term frequency *TF* is calculated based on its statistics in sliding time windows. However, the scheme [10] ignores the diversification of word frequencies in different time periods of a day. Taking words 'sleep' and 'job' as an example, Figure 6 gives the variations of tends of their frequencies at 24-hour range of a day in Australia. From Figure 1, we can see that their *TF*s have some differentiations but actually normal peaks at certain time frames. Among them, 'sleep' is talked mostly at 21:00–22:00 o'clock, while the most mentioning about 'job' appears at 17:00 o'clock. If we only use the historical data in a number of past hours for extracting bursting words, we would wrongly consider them as bursty when their normal peaks appear. Based on our idea of STWS, a reliable baseline for bursty word identification can be observed, which can further benefit the emerging event detection.

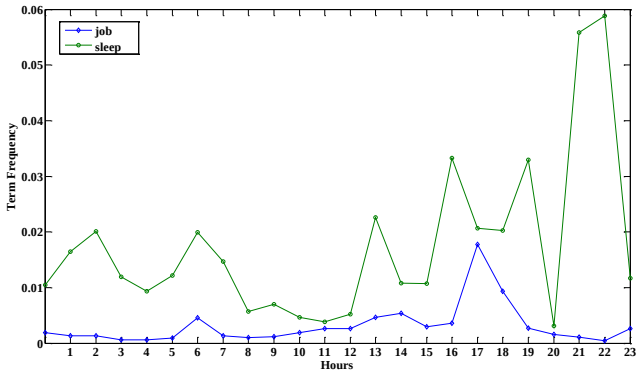

Figure 6. The *TF*s of 'sleep' and 'job'.

5.1 Predicting Events based on STWS Model

The Spatial-Temporal Word Spectrum (STWS) model provides the expected probability of a word in a given region at a certain time. Based on STWS, we can effectively extract the bursty words for a geolocation. The basic idea is to use a probabilistic approach such as the one in [10] which is based on a binomial distribution of words. Assume the normal frequency of word k at time T_i is $p_k^{(i)}$, then we can compute the probability of the number of posts consisting of word k at T_i, denoted as $P_k^{(i)}$:

$$P_k^{(i)} = \binom{N^{(i)}}{n_k^{(i)}} \left(p_k^{(i)}\right)^{n_k^{(i)}} \left(1 - p_k^{(i)}\right)^{N^{(i)} - n_k^{(i)}} \quad (6)$$

where $N^{(i)}$ is the total number of posts and $n_k^{(i)}$ is the number of posts that contains term k at time T_i. If $P_k^{(i)}$ is close or equal to zero, we take word k as a burst. Next, we use a single pass incremental clustering model [18, 19, 22] to cluster the burst-relevant posts to identify the emerging events. The performance is evaluated in following section.

5.2 Experimental Results

To evaluate the effectiveness of our STWS model for detection of the emerging local events, we use *Twitter Stream API* to have crawled 37,225 tweets in total. The tweets were collected from Sydney metropolitan area, from 25 June 2015 to 28 June 2015.

Here, we used an extended sliding window method to detect burst from data streams in our experiment. In every 10 minutes, we compare all words' frequencies in current time window (1 hour) with their baseline frequencies available from STWS. Table 2 shows two detected emerging local events or regarded as event-driven topics. We took the first event for detailed explanation as shown in Figure 7. It clearly shows the frequency variation tend of a local event related to two words: 'police' and 'drug'. It can also been seen that both words have frequencies compared to their baseline of normal circumstances, while demonstrating an obvious peak of co-occurrence at 20:00 o'clock. As we classified all the posts published during 20:00–21:00 o'clock into the data sub-set of 20:00 o'clock in experimental statistics, the peaks appear at 20:00 o'clock in Figure 2. Actually, they are detected at about 20:50 o'clock. Through a verification on Twitter and searching in our dataset, we found that there were a number of people who were complaining NSW orange polices green-lighting drug deal during this period. Therefore, we verified that our detection result is corresponding to the real-world event. In addition, the extended sliding window model can effectively improve the real-time detection of emerging local events.

Table 2 Examples of events detected by our approach.

Date	Event Topics	Description
25/06/2015	police, drug, Orange, light, sell, dealer	People dissatisfied with Orange polices' green-lighting for drug dealers
27/06/2015	marriage, Australia, legalise, gay, lovewin, equality	Appealing for same sex marriage legitimation in Australia

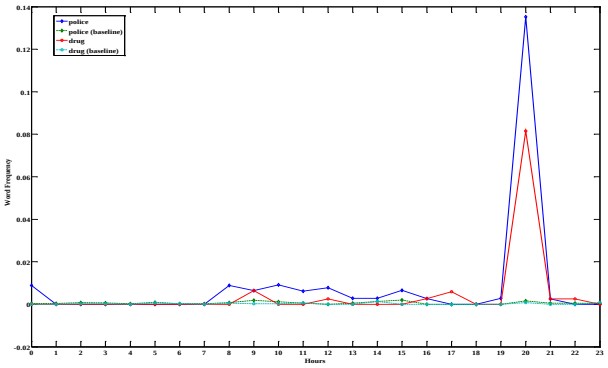

Figure 7 Frequency comparison with the baseline.

6. APPLICATION FOR GEOLOCATION PREDICATION

Geolocation predication of social media is a hot issue recently. One of the effective detection methods is based on the content analysis [3]. Because the length of a tweet is limited, the tweets were assembled by grid or by cities to enhance the performance. For example, a grid centroids method is applied on Twitter data sets from US to achieve state-of-the-art geolocation predication results [18].

6.1 K-Nearest Neighbour Using STWS Model

A K-nearest neighbour (K-NN) classifiers is commonly used in signal processing because it can achieve a good performance [20]. A K-NN classifier is applied to geolocation predication. The main idea is that given a new word spectrum sets *t*, the algorithm obtains the *K* nearest neighbouring word spectrum from the training set *Y* based on the distance between *t* and *Y* as shown in Equation (7).

$$dst(x_t, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_t)^2} \quad (7)$$

The most dominating postcode and regions amongst these K neighbours is assigned as the class of *t*. K-NN algorithm is implemented by R package FNN in this study (http://cran.r-project.org/web/packages/FNN/index.html), where *K* is chosen as 3.

6.2 Experimental Evaluation

We show how STWS is used to improve the performance of the geolocation predication. Firstly, the dataset Tweet-AU (see Section 4.1) is used as a benchmark. We extracted an STWS model for Australian states of NSW and QLD and selected the keywords shown in Table 1 (see Section 4.2). There are 12,500 tweets from 1,032 postcode areas. Secondly, the benchmark was divided into two partitions, 50% of tweets used for training and 50% of tweets used for testing. Because the areas given by postcodes are generally smaller than the area of a city, the adjacent postcodes of spectrums were clustered together for the simplification of experiments. Then a K-NN classifier was used for evaluation of the geolocation detection compared to the approached used in [1]. The results are shown in Table 3.

Table 3. Accuracy comparison of geolocation predications.

Method	22 regions	103 areas
TF* ICF [1]	36.63%	10.46%
Our approach (STWS)	38.23%	11.96%

It is clearly seen in Table 3 that the accuracy of geolocation predictions based on our spatial-temporal word spectrum model is better than the approach that purely considers the term frequencies of location-indicative words.

7. CONCLUSION

Currently, work on identifying the linguistic characteristics or 'fingerprints' of social media is still in its infancy. The Spatial-Temporal Word Spectrum (STWS) model proposed in this paper is a baseline representation for effective social media analysis. It is basically a statistical model for representing **where, when**, and **what** are available on social media over certain time periods. When large volumes of social media data are regularly collected with time intervals for different geolocations, we would be able to

see a **numerical landscape of the entire social media** for the unique representation of their spatial and temporal features.

In this paper, we illustrated the construction, visualization, and the applications of the STWS model. Particularly we demonstrated the applications of STWS model for the prediction of emerging local events and the prediction of geolocations of tweets.

We are working on the further exploration of this model for its formalization of the properties such as the convergence of a set of keywords that could be found and used as a signature of the geolocations. From a viewpoint of computational linguistics [7, 8], a complete STWS model is being built for all Australian regions.

8. ACKNOWLEDGMENT

This work is partially supported by Australian Research Council Discovery Project with Project ID ARC DP140100104.

9. REFERENCES

[1] Aggarwal C. C. & Subbian K. (2012). Event detection in social streams. In SDM, 624–635, 2012.

[2] Andrienko, N. & Andrienko, G. (2006). Exploratory analysis of spatial and temporal data: a systematic approach. Springer Science & Business Media.

[3] Bo, H. & Baldwin, P. C. T. (2012). Geolocation prediction in social media data by finding location indicative words. Proceedings of COLING 2012: Technical Papers: 1045-1062.

[4] Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis, IEEE Intelligent Systems, vol.28, no. 2, pp. 15-21.

[5] Cambria, E., Schuller, B., Liu, B., Wang, H. & Havasi, C. (2013). Knowledge-based approaches to concept-level sentiment analysis. IEEE Intelligent Systems, (2), 12-14.

[6] Cambria, E., Wang, H. & White, B. (2014). Guest editorial: Big social data analysis. Knowledge-Based Systems, (69), 1-2.

[7] Church, K. W. & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. Computational linguistics, 19(1), 1-24.

[8] Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. Applied linguistics, 25(4), 431-447.

[9] Earle, P. S., *et al.* (2012). Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics 54(6).

[10] Fung, G. P. C., *et al.* (2005). Parameter free bursty events detection in text streams. Proceedings of the 31st VLDB Conference, 181-192, 2005.

[11] Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. (2003). KNN model-based approach in classification. In On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (986-996). Springer Berlin Heidelberg.

[12] Harrison, C. (2015). Word spectrum: Visualizing Google's BiGram Data. http://www.chrisharrison.net/index.php/Visualizations/WordSpectrum.

[13] Google n-gram (2006). http://googleresearch.blogspot.com.au/2006/08/all-our-n-gram-are-belong-to-you.html.

[14] Jalal, M., *et al.* (2014). Home location identification of Twitter users. ACM Trans. Intell. Syst. Technol. 5(3): 1-21.

[15] Li, L., Goodchild, M. F. & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and geographic information science, 40(2), 61-77.

[16] Liu, B. (2015). Sentiment analysis mining opinions, sentiments, and emotions, Cambridge University Press.

[17] Stoica, P. & Moses, R. L. (2005). Spectral analysis of signals, Pearson/Prentice Hall Upper Saddle River, NJ.

[18] Stephen, R., *et al.* (2012). Supervised text-based geolocation using language models on an adaptive grid. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

[19] Unankard, S., Li, X. & Sharaf, M. A. (2014). Emerging event detection in social networks with location sensitivity. World Wide Web Journal (WWWJ), 1-25.

[20] Unankard, S., Li, X., Sharaf, M. A., Zhong, J. & Li, X. (2014). Predicting elections from social networks based on sub-event detection and sentiment analysis, In WISE, (Web Information System Engineering), Part II, LNCS8787, 1-16.

[21] Unankard, S., Li, X. & Long, G. (2015). Invariant event tracking on social networks. In Database Systems for Advanced Applications, DASFAA2015, 517-521, 2015.

[22] Yin, J., *et al.* (2012). Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 52-59, 2012.

[23] Zhu, G., *et al.* (2014). "Epileptic seizure detection in EEGs signals using a fast weighted horizontal visibility algorithm." Computer Methods and Programs in Biomedicine 115(2): 64-75.