Stopword Detection for Streaming Content

Hossein Fani[1,3], Masoud Bashari[1,2], Fattane Zarrinkalam[1,2], Ebrahim Bagheri[1], and Feras Al-Obeidat[1]

[1] Laboratory for Systems, Software and Semantics (LS[3]), Ryerson University
[2] Dept. of Computer Engineering, Ferdowsi University of Mashhad
[3] Faculty of Computer Science, University of New Brunswick

Abstract. The removal of stopwords is an important preprocessing step in many natural language processing tasks, which can lead to enhanced performance and execution time. Many existing methods either rely on a predefined list of stopwords or compute word significance based on metrics such as tf-idf. The objective of our work in this paper is to identify stopwords, in an unsupervised way, for streaming textual corpora such as Twitter, which have a temporal nature. We propose to consider and model the dynamics of a word within the streaming corpus to identify the ones that are less likely to be informative or discriminative. Our work is based on the discrete wavelet transform (DWT) of word signals in order to extract two features, namely *scale* and *energy*. We show that our proposed approach is effective in identifying stopwords and improves the quality of topics in the task of topic detection.

1 Introduction

Stopwords are non informative or noise tokens which can and should be removed during any preprocessing step. Removing stopwords not only decreases computational complexity [15] but can also improve the quality of the final output in document processing tasks such as clustering [11], indexing [14], and event detection [8]. There is existing work that focus on removing task-specific non-informative words from a collection of documents [10, 13]. In these approaches, words are assumed to be features and off-the-shelf feature selection methods such as mutual information are used to identify the most important features (words) with respect to the target task. As a result, in the same text corpus, the list of unimportant words depends on the target task and is not reusable in other tasks. On the contrary, our goal is to devise a systematic procedure to identify an effective list of stopwords which would be agnostic to the underlying task. In this vein, Salton et al. [14] proposed the tf-idf word weighting scheme in information retrieval and document indexing. According to the tf-idf scheme, a stopword's *average* tf-idf value over all documents is either very high or very low. The latter means that the word is common to all documents and, consequently cannot discriminate documents from each other while the former indicates that the word is rare and does not contribute significantly to the corpus. Tf-idf has seen many variations since its inception, e.g., term frequency-proportional document frequency (tf-pdf) [4] and has been applied in different domains and languages such as [3, 5, 13], but none has incorporated the temporal stream-based (real-time) nature of textual collections in online social networks to identify stopwords.

Online social content are inherently temporal, short and noisy about transient topics that are generated in bursts by users and fade away shortly afterwards. It is in contrast to long formal documents that are often well structured and the whole corpus is available in advance. Researchers have looked at ways through which word significance can be computed so as to dynamically identify stopwords in different periods of time. For instance, Weng et al. [15] and Fani et al. [7] employ signal cross-correlation to filter out trivial words. He et al. [8] use spectral analysis and perform word feature clustering for the same purpose. Both methods distinguish themselves from the traditional way of stopword detection by taking the temporal dimension into account. Inspired from these works, in this paper, we propose a discrete wavelet transform (DWT) of word signals for identifying stopwords and explore its effectiveness in the context of topic detection on Twitter.

To the best of our knowledge, the two work presented in [7, 8] are the state of the art in identifying stopwords based on the temporal dynamics of words in a temporal corpus and as such they comprise our baselines with which we evaluate our proposed approach. We also include tf-idf to our baselines as the most common domain independent best practice for finding stopwords.

2 Proposed Approach

The main objective of our work is to identify stopwords in a temporal stream of textual content. To this end, we first build word signals based on their occurrences sampled across time intervals as follows:

Definition 1. ***(Word Signal)*** *The signal of a word w is a sequence $X^w = [x^w_{1:T}]$ where each element x^w_t shows the number of times w occurred in time interval t.*

We transform the word signals from the time domain to the time-scale (frequency) domains by discrete wavelet transform (DWT) in order to extract two features, namely *scale* and *energy*. This is similar to the spectral analysis done by He et al. [8] where Fourier transform is used to extract power and periodicity features from word signals. However, while Fourier transform only informs us about the frequency spectrum of the signal, we opt for wavelet transform for its ability to disclose temporal extent of the signal (time) in addition to the frequency spectrum (scale is inversely proportional to frequency, i.e., scale $\propto \frac{1}{freq.}$). Fourier transform does not reveal bursty changes efficiently since it represents a signal as the sum of sine waves, which are not localized in time. To accurately analyze signals with rich dynamics in time and frequency, wavelets are used. A wavelet is a rapidly decaying, wave-like oscillation that has zero mean. Unlike sinusoids, which extend to infinity, a wavelet exists for a finite duration. The foundational representation of the discrete wavelet transform is [9]:

$$W_\Psi\{X^w\}(a,b) = \frac{1}{\sqrt{2^a}} \sum_{t=0}^{T} x^w_t \Psi(t - 2^a b) \tag{1}$$

where x^w_t is the value of word signal at time t, Ψ is the base (mother) wavelet, and 2^a and b define the scale and shifting (translation) in dyadic sampling, respectively with

$a \in \mathbb{Z}^+, b \in \mathbb{Z}$. This kind of sampling eliminates redundancy in coefficients and yields the same number of wavelet coefficients as the length of the input word signal, i.e., T, for *all* scales. Otherwise, in normal sampling the number of wavelet coefficients would be T for *each* scale. Also Ψ is chosen to be Mexican-hat for its high temporal resolution.

To figure out which scales play a more important role to reconstruct the entire original word signal, we calculate the total energy at each scale, which is a sum of wavelet coefficients over the whole signal through shifting variable b, as follows [9]:

$$E_\Psi\{X^w\}(a) = \sum_b |W_\Psi\{X^w\}(a,b)|^2 \tag{2}$$

Next, given energy of word signals in different scales, we classify the words into four categories: (1) *Significant* words are such words that have signals of high energy in low scales (high frequency); (2) *Common* words' are such words that have signals of high level of energy in high scales (low frequency). This means they have approximately constant behaviour with high amplitude over different time intervals; (3) *Noise* words have a low occurrence and hence low energy in low scales (high frequency); (4) *Rare* words denote the words that have low energy in high scales (low frequency).

3 Experiments

We evaluate our wavelet-based approach (Wvl) against three baselines; Fourier-based (Ftp) [8], cross-correlation (Xcr) [7], average tf-idf, and random selection in finding high quality emerging topics using latent Dirichlet allocation (LDA) topic modelling. We hypothesize that stopword removal produce topics with higher quality in the task of topic detection. Conversely, we conclude that higher quality topics imply better stopword removal assuming the dataset and the topic detection method are both kept constant in our experiments.

To evaluate the performance of the proposed wavelet-based (Wvl) approach and the baselines, we first identify the stopwords by each method for Abel et al.'s Twitter dataset [1]. This dataset consists of 3M tweets posted by 135,731 unique users between Nov. 1 and Dec. 31, 2010. In each approach, we create a list of all words ordered by their significance score and incrementally build the stopword list of from the least to the most significant words. This has been initiated from the empty list of size zero, i.e., no stopwords, increasing by 10% at each iteration to 90% of all words. Also all temporal approaches, i.e., Wvl, Ftp, and Xcr, are examined using both hourly and daily sampling rates.

We then apply LDA, the *de facto* standard in topic modeling,to detect emerging topics using mallet[4] after removing the stopwords. The number of topics has been already investigated and set to 50 for the this dataset [6]. Finally, we use *exclusivity*, *specificity*, and *coherence* as quality metrics for the output topics. It should be noted

[4] mallet.cs.umass.edu/topics.php

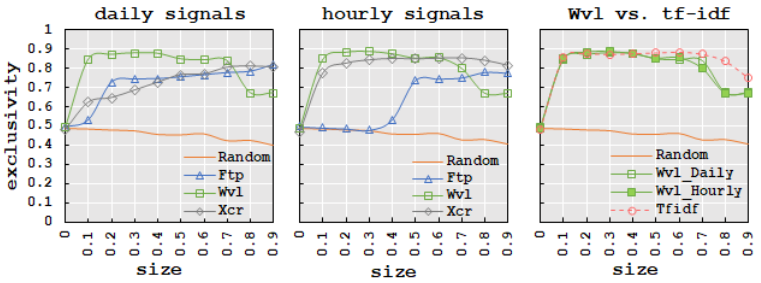

Fig. 1: Average exclusivity of topics.

that perplexity and the likes which measure the quality of the topic modeling approaches, not the topics, cannot be applied here. In the following, we explain each metric followed by a discussion on the respective results.

Exclusivity [2] measures to what extent the top words for a topic do not appear as top words in other topics, as follows:

$$e(z) = \frac{1}{|z|} \sum_{i=1}^{|z|} \frac{Pr(w_i|z)}{\sum_{z' \neq z} Pr(w_i|z')} \tag{3}$$

where z is a topic consisting of a list of |z|=50 words ordered by probability of occurring in the topic and $P(w|z)$ is the probability of word w occurring in topic z.

Figure 1 shows the exclusivity values under varying size of stopword lists. The figure reveals that our proposed wavelet-based (Wvl) approach reaches the highest exclusivity while removing less words (10%) in both daily and hourly signals indicating best stopword removal among all the temporal competitors. Further, in contrast to Wvl, Fourier-based (Ftp) and cross-correlation (Xcr), deliver high sensitivity to the signal sampling rate. As shown in Figure 1, Ftp in daily signals outperforms its hourly version as it reaches its peak sooner where 20% of the words are removed as stopwords. However, Xcr excels in hourly signals compared to its daily version. This points to the robustness of the proposed Wvl approach with regards to sampling rate. By comparing average tf-idf (non-temporal) and Wvl (best temporal), exclusivity scores are very competitive.

Specificity[5] measures the effective number of words in each topic. The idea is to count topic's words while applying the probability that a word is contributing to a topic. Higher specificity implies that words are distributed in different topics exclusively and are not conforming to the uniform distribution. Specificity is calculated as:

$$s(z) = \sum_{w \in z} \frac{1}{Pr^2(w|z)} \tag{4}$$

In Figure 2, we show the average specificity values over all topics for the proposed Wvl approach and the baselines. As evident, the Wvl approach significantly excels the specificity score both in daily and hourly signals when removing 20% of words

[5] mallet.cs.umass.edu/diagnostics.php

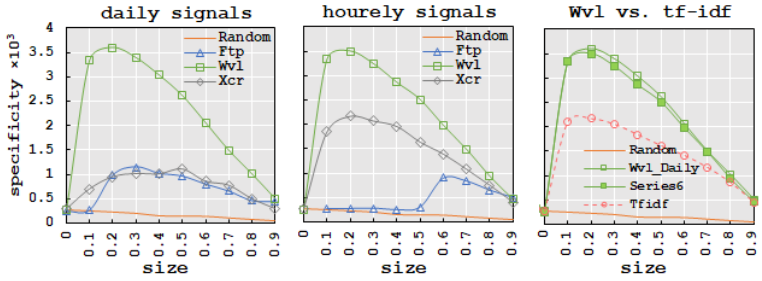

Fig. 2: Average specificity of topics.

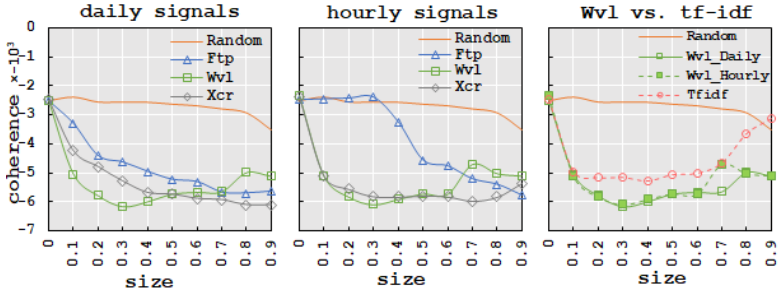

Fig. 3: Average coherence of topics.

as stopwords in comparison to not only the temporal baselines but also the average tf-idf.

Coherence [12] measures whether the words in a topic tend to co-occur together. The score is the sum of log probability that a document contains at least one instance of a the higher-ranked and a lower-ranked word pair. Formally,

$$c(z) = \sum_{i=1}^{|z|} \sum_{j<i} log \frac{D(w_i, w_j) + \epsilon}{D(w_i) + \epsilon} \tag{5}$$

where ϵ is a smoothing parameter set to 0.01, $D(w)$ is the number of documents that contain at least one token of type w, $D(w_i, w_j)$ is the number of documents that contain at least one w_i and one w_j. Since these scores are log probabilities, they are negative. High values (closer to zero) indicate that words of the topic tend to co-occur more often; hence, the topic is coherent. Figure 3 shows the average coherence values for the identified LDA topics. A first look reveals that our proposed Wvl is the worst and random selection of words as stopwords outperforms all the other approaches. As such, we manually inspect this seemingly unintuitive phenomenon and found out that stopwords which are common to all documents such as 'rt', 'http', and 'us' contribute to almost all topics with a high probability (top words of almost all topics). Removing such words greatly reduces coherence as shown in Figure 3. In other words, in the task of stopword removal, the lower coherence to an extremum the better. With this respect, our proposed Wvl is the best and the hourly cross-correlation is the runner up.

4 Concluding Remarks

In this paper, we propose a wavelet-based approach for identifying stopwords from streaming textual corpora with temporal nature in an unsupervised way. We employ discrete wavelet transform to model the dynamics of a word within the streaming corpus. We identify informative words based on their energy over different scales of word signals. We showed that our proposed approach is able to improve *exclusivity* and *specificity* of topics learnt based on LDA when the identified stopwords were removed from the corpus. Further, we observed that the current definition of the *coherence* metric is ineffective for the task of stopword removal. Common words happen to be in the top words of almost all topics, the removal of which drops *coherence* significantly. For future work, we will investigate the potential performance improvements of our proposed method in the context of higher-level applications such as user interest modeling, news recommendation and community detection.

References

1. F. Abel, Q. Gao, G. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *UMAP'11*.
2. J. M. Bischof and E. M. Airoldi. Summarizing topical content with word frequency and exclusivity. In *ICML'12*, pages 9–16, USA, 2012. Omnipress.
3. A. Blanchard. Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308, Dec. 2007.
4. K. K. Bun and M. Ishizuka. Emerging topic tracking system in WWW. *Knowl.-Based Syst.*, 19(3):164–171, 2006.
5. K. Darwish, W. Magdy, and A. Mourad. Language processing for arabic microblog retrieval. In *CIKM'12*, pages 2427–2430, 2012.
6. H. Fani, E. Bagheri, F. Zarrinkalam, X. Zhao, and W. Du. Finding diachronic like-minded users. *Computational Intelligence: An International Journal*, 2017.
7. H. Fani, F. Zarrinkalam, X. Zhao, Y. Feng, E. Bagheri, and W. Du. Temporal identification of latent communities on twitter. *CoRR*, abs/1509.04227, 2015.
8. Q. He, K. Chang, and E. Lim. Analyzing feature trajectories for event detection. In *SIGIR'07*, pages 207–214, 2007.
9. G. Kaiser. *A Friendly Guide to Wavelets*. Birkhauser Boston Inc., Cambridge, MA, USA, 1994.
10. B. Klatt, K. Krogmann, and V. Kuttruff. Developing stop word lists for natural language program analysis. *Softwaretechnik-Trends*, 34(2), 2014.
11. X. Li, J. Chen, and O. R. Zaïane. Text document topical recursive clustering and automatic labeling of a hierarchy of document clusters. In *PAKDD*, pages 197–208, 2013.
12. D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272, 2011.
13. S. Popova, T. Krivosheeva, and M. Korenevsky. Automatic stop list generation for clustering recognition results of call center recordings. In *Speech and Computer - 16th International Conference, SPECOM 2014*, pages 137–144, 2014.
14. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.
15. J. Weng and B. Lee. Event detection in twitter. In *ICWSM'11*, 2011.