Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter

Hassan Saif, Miriam Fernandez and Harith Alani

Knowledge Media Institute, The Open University, United Kingdom
{h.saif,m.fernandez,h.alani}@open.ac.uk

Abstract. In this paper we propose a semantic approach to automatically identify and remove stopwords from Twitter data. Unlike most existing approaches, which rely on outdated and context-insensitive stopword lists, our proposed approach considers the contextual semantics and sentiment of words in order to measure their discrimination power. Evaluation results on 6 Twitter datasets show that, removing our semantically identified stopwords from tweets, increases the binary sentiment classification performance over the classic pre-complied stopword list by 0.42% and 0.94% in accuracy and F-measure respectively. Also, our approach reduces the sentiment classifier's feature space by 48.34% and the dataset sparsity by 1.17%, on average, compared to the classic method.

Keywords: Sentiment Analysis, Contextual Semantics, Stopwords, Twitter

1 Introduction

The excessive presence of abbreviations and irregular words in tweets make them very noisy, sparse and hard to extract sentiment from [7, 8]. Aiming to address this problem, existing works on Twitter sentiment analysis remove stopwords from tweets as a pre-processing procedure [5]. To this end, these works usually use pre-complied lists of stopwords, such as the *Van stoplist* [3]. These stoplists, although widely used, have previously been criticised for: (i) being outdated [2] and, (ii) for not accounting for the specificities of the context under analysis [1]. Words with low informative values in some context or corpus, may have discrimination power in a different context. For example, the word "`like`", generally consider as stopword, has an important sentiment discrimination power in the sentence "`I like you`".

In this paper, we propose an unsupervised approach for automatically generating context-aware stoplists for the sentiment analysis task on Twitter. Our approach captures the contextual semantics and sentiment of words in tweets in order to calculate their informative value. Words with low informative value are then selected as stopwords. Contextual semantics (aka statistical semantics) are based on the proposition that meaning can be extracted from words co-occurrences [9].

We evaluate our approach against the *Van stoplist* (so-called clasic method) using six Twitter datasets. In particular, we study how removing stopwords generated by our approach affects: (i) the level of data sparsity of the used datasets and (ii) the performance of the Maximum Entropy (MaxEnt) classifier in terms of: (a) the size of the classifier's feature space and, (b) the classifier's performance. Our preliminary results show that our approach outperforms the classic stopword removal method in both accuracy and F1-measure by 0.42% and 0.94% respectively. Moreover, removing our semantically-identified stopwords reduces the feature space by 48.34% and the dataset sparsity by 1.17%, compared to the classic method, on average.

2 Stopwords Generation using Contextual Semantics

The main principle behind our approach is that the informativeness of words in sentiment analysis relies on their semantics and sentiment within the contexts they occur. Stopwords correspond to those words of weak contextual semantics and sentiment.Therefore, our approach functions by first capturing the contextual semantics and sentiment of words and then calculating their informative values accordingly.

2.1 Capturing Contextual Semantics and Sentiment

To capture the contextual semantics and sentiment of words, we use our previously proposed semantic representation model SentiCircles [6].

In summary, the SentiCircle model extracts the contextual semantics of a word from its co-occurrences with other words in a given tweet corpus. These co-occurrences are then represented as a geometric circle which is subsequently used to compute the contextual sentiment of the word by applying simple trigonometric identities on it. In particular, for each unique term m in a tweet collection, we build a two-dimensional geometric circle, where the term m is situated in the centre of the circle, and each point around it represents a context term c_i (i.e., a term that occurs with m in the same context). The position of c_i, as illustrated in Figure 1, is defined jointly by its Cartesian coordinates x_i, y_i as:

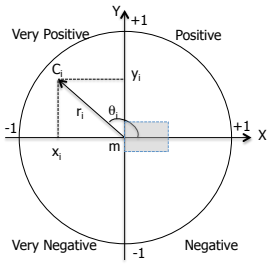

Fig. 1: SentiCircle of a term m. Stopwords region is shaded in gray.

$$x_i = r_i \cos(\theta_i * \pi) \qquad\qquad y_i = r_i \sin(\theta_i * \pi)$$

Where θ_i is the polar angle of the context term c_i and its value equals to the prior sentiment of c_i in a sentiment lexicon before adaptation, r_i is the radius of c_i and its value represents the degree of correlation (tdoc) between c_i and m, and can be computed as:

$$r_i = tdoc(m, c_i) = f(c_i, m) \times \log(N/N_{c_i})$$

where $f(c_i, m)$ is the number of times c_i occurs with m in tweets, N is the total number of terms, and N_{c_i} is the total number of terms that occur with c_i. Note that all terms' radii in the SentiCircle are normalised. Also, all angles' values are in radian.

The trigonometric properties of the SentiCircle allow us to encode the contextual semantics of a term as *sentiment orientation* and *sentiment strength*. Y-axis defines the sentiment of the term, i.e., a positive y value denotes a positive sentiment and vice versa. The X-axis defines the sentiment strength of the term. The smaller the x value, the stronger the sentiment.[1] This, in turn, divides the circle into four sentiment quadrants. Terms in the two upper quadrants have a positive sentiment ($\sin\theta > 0$), with upper left quadrant representing stronger positive sentiment since it has larger angle values than those in the top right quadrant. Similarly, terms in the two lower quadrants have negative sentiment values ($\sin\theta < 0$). Moreover, a small region called the "*Neutral Region*" can be defined. This region is located very close to X-axis in the "*Positive*" and the "*Negative*" quadrants only, where terms lie in this region have very weak sentiment (i.e, $|\theta| \approx 0$).

[1] This is because $\cos\theta < 0$ for large angles.

The overall Contextual Semantics and Sentiment An effective way to compute the overall sentiment of m is by calculating the geometric median of all the points in its SentiCircle. Formally, for a given set of n points $(p_1, p_2, ..., p_n)$ in a SentiCirlce Ω, the 2D geometric median g is defined as: $g = \arg\min_{g \in \mathbb{R}^2} \sum_{i=1}^{n} ||p_i - g||_2$. We call the geometric median g the **SentiMedian** as its position in the SentiCircle determines the total contextual-sentiment orientation and strength of m.

2.2 Detecting Stopwords with SentiCircles

Stopwords in sentiment analysis are those who have weak semantics and sentiment within the context they occur. Hence, stopwords in our approach are those whose SentiMedians are located in the SentiCircle within a very small region close to the origin, as shown in Figure 1. This is because points in this region have: (i) very weak sentiment (i.e., $|\theta| \approx 0$) and (ii) low importance or low degree of correlation (i.e., $r \approx 0$). We call this region the *stopword region*. Therefore, to detect stopwords in our approach, we first build a SentiCircle for each word in the tweet corpus, calculate its overall Contextual semantics and sentiment by means of its SentiMedian, and check whether the word's SentiMedian lies within the stopword region or not.

We assume the same stopword region boundary for all SentiCircles emerging from the same Twitter corpus, or context. To compute these boundaries we first build the SentiCircle of the complete corpus by merging all SentiCircles of each individual term and then we plot the density distribution of the terms within the constructed SentiCircle. The boundaries of the stopword region are delimited by an increase/decrease in the density of terms along the X- and Y-axis. Table 1 shows the X and Y boundaries of the stopword region for all Twitter datasets that we use in this work.

Dataset	OMD	HCR	STS-Gold	SemEval	WAB	GASP
X-boundary	0.0001	0.0015	0.0015	0.002	0.0006	0.0005
Y-boundary (Y)	0.0001	0.00001	0.001	0.00001	0.0001	0.001

Table 1: Stopword region boundary for all datasets

3 Evaluation and Results

To evaluate our approach, we perform binary sentiment classification (positive / negative classification of tweets) using a MaxEnt classifier and observe fluctuations (increases and decreases) after removing stopwords on: the classification performance, measured in terms of accuracy and F-measure, the size of the classifier's feature space and the level of data sparsity. To this end, we use 6 Twitter datasets: *OMD, HCR, STS-Gold, SemEval, WAP and GASP* [4]. Our baseline for comparison is the *classic method*, which is based on removing stopwords obtained from the pre-complied *Van stoplist* [3].

Figure 2 depicts the classification performance in accuracy and F1-measure as well as the reduction in the classifier's features space obtained by applying our SentiCircle stopword removal methods on all datasets. As noted, our method outperforms the classic stopword list by 0.42% and 0.94% in accuracy and F1-measure on average respectively. Moreover, we observe that our method shrinks the feature space substantially by 48.34%, while the classic method has a reduction rate of 5.5% only.

Figure 3 shows the average impact of the SentiCircle and the classic methods on the sparsity degree of our datasets. We notice that our SentiCircle method always lowers the sparsity degree of all datasets by 1.17% on average compared to the classic method.

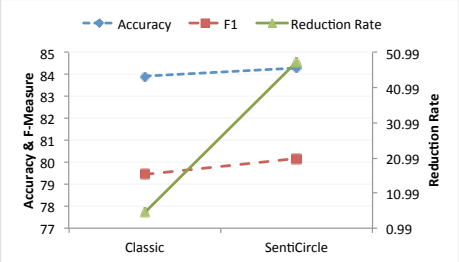

Fig. 2: Average accuracy, F-measure and reduction rate of MaxEnt using different stoplists

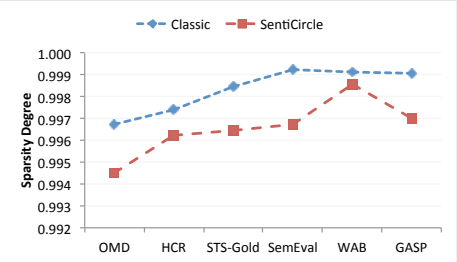

Fig. 3: Impact of the classic and SentiCircles methods on the sparsity degree of all datasets.

4 Conclusions

In this paper we proposed a novel approach for generating context-aware stopword lists for sentiment analysis on Twitter. Our approach exploits the contextual semantics of words in order to capture their context and calculates their discrimination power accordingly. We have evaluated our approach for binary sentiment classification using 6 Twitter datasets. Results show that our stopword removal approach outperforms the classic method in terms of the sentiment classification performance and the reduction in both the classifier's feature space and the dataset sparsity.

Acknowledgment

This work was supported by the EU-FP7 project SENSE4US (grant no. 611242).

References

1. Ayral, H., Yavuz, S.: An automated domain specific stop word generation method for natural language text classification. In: International Symposium on Innovations in Intelligent Systems and Applications (INISTA) (2011)
2. Lo, R.T.W., He, B., Ounis, I.: Automatically building a stopword list for an information retrieval system. In: Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR) (2005)
3. Rijsbergen, C.J.V.: Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edn. (1979)
4. Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. In: Proceedings, 1st ESSEM Workshop. Turin, Italy (2013)
5. Saif, H., Fernandez, M., He, Y., Alani, H.: On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In: Proc. 9th Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland (2014)
6. Saif, H., Fernandez, M., He, Y., Alani, H.: Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In: Proc. 11th Extended Semantic Web Conf. (ESWC). Crete, Greece (2014)
7. Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment analysis. In: Proc. 2nd Workshop on Making Sense of Microposts (#MSM2012). Layon, France (2012)
8. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proceedings of the 11th international conference on The Semantic Web. Boston, MA (2012)
9. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1), 141–188 (2010)