

Sentiment-Preserving Reduction for Social Media Analysis

Sergio Hernández¹ and Philip Sallis²

¹ Laboratorio de Procesamiento de Información Geoespacial.
Universidad Católica del Maule, Talca, Chile

² Geoinformatics Research Centre.
Auckland University of Technology, Auckland, New Zealand.

Abstract. In this paper, we address the problem of opinion analysis using a probabilistic approach to the underlying structure of different types of opinions or sentiments around a certain object. In our approach, an opinion is partitioned according to whether there is a direct relevance to a latent topic or sentiment. Opinions are then expressed as a mixture of sentiment-related parameters and the noise is regarded as data stream errors or spam. We propose an entropy-based approach using a value-weighted matrix for word relevance matching which is also used to compute document scores. By using a bootstrap technique with sampling proportions given by the word scores, we show that a lower dimensionality matrix can be achieved. The resulting noise-reduced data is regarded as a sentiment-preserving reduction layer, where terms of direct relevance to the initial parameter values are stored

1 Introduction

Social networks have become ubiquitous and are used throughout the world for interpersonal communication. This form of discourse can be related to personal matters but is also about common interests, especially products and services. In particular these discussions about products and services give rise to a massive source of valuable information. In this paper, focused as it is on text communications on the web, it centers on online discussions, tweets and social networks, which are subjective in nature and therefore not easy to classify. Mostly, these communications/conversations concern the quality of a particular product.

Sentiment analysis is used in information retrieval and text mining for discovering the attitude or the subjective judgment of the writer about a particular matter. For example, it is used in social media to make judgments about certain products and services that are of interest to them. The massive amount of data that is generated by this media can be used to optimize the commodities, by analyzing the overall sentiment towards expressed about them [8].

In general, sentiment can be expressed as a quantity (e.g. a score) or as a textual opinion. The latter might reveal a polarity that can be unveiled by using data analytics [4]. In the context of data mining and knowledge discovery we can distinguish two main approaches, supervised and unsupervised learning, as being

possible ways to recognize patterns in the text based discourse being analyzed. Sentiment analysis or hidden meaning can be regarded as being unsupervised, because when triggered by an associated polarity, the hidden meaning is revealed.

Likes and dislikes are often concealed in words where the meaning less than obvious and in some cases, criticism can be encapsulated in a cynical or somewhat ambiguous phrase. Linguistic markers such as the identification of particular words from a list or table on the assumption that may mean something in particular could lead to a completely different meaning from the one intended [5].

For example, when analyzing sentiments about the iPad©we can find phrases containing a sentiment, such as:

```
I can't believe how fast twitter works on my iPad
```

In the other hand, phrases like

```
Free iPad 2! How awesome is that? You HAVE to join!
```

doesn't contain any sentiment information so they can be considered as spam. Moreover, we are also confronted with the reality that individual words, while having some meaning, are not richly enough preserved until they appear alongside other words in terms and phrases. In our approach, there is no prior knowledge of the sentiment or polarity of a phrase. Alternatively, an entropy-based criteria based on a non-uniform prior distribution on words can be used to gather sentiment-preserving information. We present a worked example using a *bag-of-words* representation, but the proposed approach can be generalized to other representations based on multiple correlated words [1].

This paper is organized as follows. Section 2 discusses the general framework for sentiment analysis and similar approaches in the literature. In Section 3, we describe the proposed methodology for sentiment-preserving reduction and finally, Section 4 provides a worked example using Twitter.³

2 Latent Topic Opinion Mining

Probabilistic models such as *topic models* can be used to discover the hidden or latent description or the topic of a group of opinions using a particular combination of words [11]. In topic modeling, a *document-term matrix* X is extracted from a text corpora. This matrix describes the occurrences of terms in documents and is composed the frequency on each one of the phrases, so each element x_{ij} contains the frequency of the word w_i in the document or opinion o_j .

A probabilistic model could consider each word as a mixture of single or multiple words (n-grams) and each opinion o being generated by first choosing a topic z and then sampling N words according to the conditional distribution $p(o)$ of words given the topic:

³ <http://twitter.com>

$$p(o) = \sum_z p(z) \prod_{n=1}^N (p(w_n|z)) \quad (1)$$

If we now let each opinion to exhibit not only one but multiple topics (e.g. words having more than one meaning), the resulting generative model for a word is a mixture of multinomial random variables representing the different topics.

$$p(o, w_n) = p(o) \sum_z p(z) p(w_n|z) p(z|o) \quad (2)$$

Each opinion is then represented as a list of mixture proportions representing its membership to any particular topic. Due to the bag-of-words assumption, there is no particular order for the words w_n so the probabilistic approach is simplified. However, the frequency of counts approach might not be enough to capture the structure of the opinions and because of the large number of parameters required is also likely to pose over-fitting issues. This is especially problematic in opinion mining where the number of number of words is usually smaller than the standard documents considered in topic modeling.

Latent Dirichlet Allocation (LDA) extends the probabilistic approach based on mixtures of unigrams by considering exchangeable partition of the set $\{z_1, \dots, z_N\}$. In LDA, words are generated by conditionally independent and identically distributed topics, so the probability of a sequence of words and topics can be written as the product:

$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^N p(z_n|\theta) p(w_n|z_n) \right) d\theta \quad (3)$$

The parameter θ is used for the multinomial distribution for each topic. Now, using Dirichlet prior distributions with hyper-parameters α and β for the topics and words respectively, leaves the following generative process:

```

Choose  $\theta \sim Dir(\alpha)$ 
for  $n = 1$  TO  $N$  do
    Choose a topic  $z_n \sim M(\theta)$ 
    Choose a word  $w_n$  from the conditional distribution of the word given the
    chosen topic  $p(w_n|z_n, \beta)$ 
end for

```

2.1 Related work

A Joint Sentiment/Topic (JST) model was proposed in [7]. In their approach, sentiment polarity is treated as an unsupervised learning problem where sentiment and topic are jointly detected from text using LDA. Given the fact that

sentiment can be expressed in a more subtle way than a topic, the authors proposed to incorporate prior information by using a subjectivity lexicon with aggregated words displaying positive or negative polarity.

More recently, the JST approach has been extended into a weakly-supervised approach in [6], and the authors reported improved classification accuracy when compared to semi-supervised alternatives. Also, in the context of micro-blogging, qualitative and quantitative experiments on topic modeling using short text messages was studied in [3].

More closely related to this work, a sentiment-preserving dimension reduction methodology has been presented in [10]. The authors proposed an inverse projection from word frequencies into sentiment, where prior knowledge of the conditional distribution is used for the inverse regression of text. This approach requires labeled data and was tested in richer text corpora, such as political speeches and restaurant reviews. Instead, in our approach we analyze data from micro-blogging environments which is not labeled, so there is no prior knowledge of the sentiment of the documents. A previous article also describes the proposed methodology [2].

3 Sentiment Preserving Reduction

In order to perform opinion spam detection we would like to find a matrix \hat{X} with lower dimensionality than the original matrix X . A signal denoising algorithm based on entropy can be then used to eliminate columns with non useful phrases leaving only text meaning vectors. This requires us to process a large quantity of data in order to identify errors in the signal stream and thereby, generate the matrix of non-noisy items.

For a particular set of M opinions and N_d words, the entropy is given by:

$$p(O_M) = \exp\left(-\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d}\right) \quad (4)$$

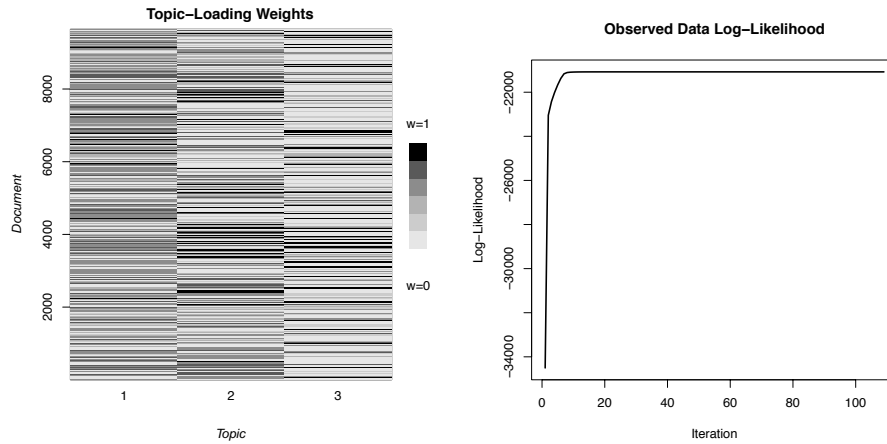
Because the number of opinions M is usually large, a brute force implementation for spam detection is not feasible. However, we can take a sample from a bootstrap sample and then compare the information gain from the entropy of the LDA model having a term matrix X_{test} . This procedure can be repeated until some criteria of convergence is achieved.

The following algorithm shows this methodology:

```

repeat
  Find a subset  $O_J \subset O_M$  with  $J < M$ .
  if  $P(O_J) < P(O_M)$  then
    Let  $O_M = O_J$ 
  end if
until Convergence

```

(a) Topic distribution for the complete model

(b) Log-likelihood

Fig. 2: Posterior distribution of the complete model

Topic 1	Topic 2	Topic 3
app	free	iphone
apple	apple	ipod
apps	win	join
tablet	amp	link
via	copy	click

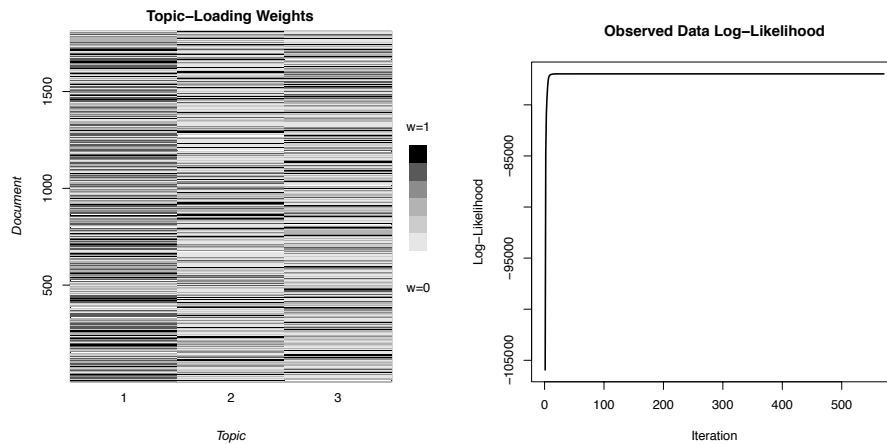
Table 1: Top five terms in the complete model

tribution for the sentiment-preserving dimensionality reduction algorithm. The following terms were considered as relevant to the sentiment analysis task:

"love", "free", "waste", "compatibility", "compatibility", "cheesy", "great", "obscure", "fantastic", "low", "fine", "cost", "speed"

Using the sentiment-preserving algorithm, a vector-valued weight function is then applied to sample a portion of the original dataset that has better entropy than the complete dataset. The posterior probabilities using the sentiment-preserving reduction are shown in Figure 3.

The five most frequent terms of the reduced model are also shown in Table 2 and the summary statistics of each model is shown in Table 3



(a) Topic distribution for the reduced model

(b) Log-likelihood

Fig. 3: Posterior distribution of the reduced model

Topic 1	Topic 2	Topic 3
handing	ipad	ipad
broken	tweets	giving
bugging	increase	entry
howdy	app	win
giving	free	free

Table 2: Top 5 terms in the reduced model

5 Conclusion

Opinions are usually populated with words and phrases having subjective meanings. Probabilistic topic models can represent sentiment in opinions by modeling the uncertainty of words and topics. In this regard, sentiment becomes a signal yet to be discovered through multiple and hidden topics. However, the amount of spam in social media can lead to deceitful results.

Here, we have presented an unsupervised sentiment-preserving data reduction method. The method is based on the standard Latent Dirichlet Allocation methodology; thus not requiring any classification of polarity in the opinions. Similar to the previously proposed Joint Sentiment-Topic model, our method is also based on a manually selected subjectivity lexicon. However, we only use it as a proxy to a bootstrapping technique that gathers sentiment-rich opinions. We have demonstrated that the resulting reduction has better entropy than the model using the complete dataset, indicating better generalization performance.

	complete model	reduced model
Entropy	44.56	33.32
# of documents	9643	1811
Sparsity	95%	96%

Table 3: Summary statistics of the complete and the reduced model

Since our method is completely unsupervised, there is no direct interpretation of the sentiment over topics. Further research in semi-supervised and instrumental regression techniques will be conducted for the sentiment detection problem.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (March 2003), <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>
2. Hernandez, S., Garden, K.L., Sallis, P.J.: A signal denoising method for text meaning vectors. In: *Proceedings of the Fifth Asia Modelling Symposium (To Appear)* (2011)
3. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: *Proceedings of the First Workshop on Social Media Analytics*. pp. 80–88. SOMA '10, ACM, New York, NY, USA (2010), <http://doi.acm.org/10.1145/1964858.1964870>
4. Hu, M., Liu, B.: Opinion extraction and summarization on the web. In: *proceedings of the 21st national conference on Artificial intelligence - Volume 2*. pp. 1621–1624. AAAI Press (2006), <http://portal.acm.org/citation.cfm?id=1597348.1597456>
5. Jindal, N., Liu, B.: Opinion spam and analysis. In: *Proceedings of the international conference on Web search and web data mining*. pp. 219–230. WSDM '08, ACM, New York, NY, USA (2008), <http://doi.acm.org/10.1145/1341531.1341560>
6. Lin, C., He, Y., Everson, R., Ruger, S.: Weakly-supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on PP(99)*, 1 (2011)
7. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: *Proceeding of the 18th ACM conference on Information and knowledge management*. pp. 375–384. CIKM '09, ACM, New York, NY, USA (2009), <http://doi.acm.org/10.1145/1645953.1646003>
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* 2, 1–135 (January 2008), <http://portal.acm.org/citation.cfm?id=1454711.1454712>
9. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2011), <http://www.R-project.org/>, ISBN 3-900051-07-0
10. Taddy, M.A.: *Inverse Regression for Analysis of Sentiment in Text*. ArXiv e-prints (Dec 2010)
11. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: *Proceedings of the 23rd international conference on Machine learning*. pp. 977–984. ICML '06, ACM, New York, NY, USA (2006), <http://doi.acm.org/10.1145/1143844.1143967>