

A Signal Denoising Method for Text Meaning Vectors

Sergio Hernández
Lab. de Proc. de Inf. Geoespacial
Universidad Católica del Maule
Talca, Chile
shernandez@ucm.cl

Philip Sallis
Geoinformatics Research Centre
Auckland University of Technology
Auckland, New Zealand
philip.sallis@aut.ac.nz

Kathy Garden
Information Technology Research Ltd
Auckland, New Zealand
kathy@itrl.co.nz

Abstract—The extraction of meaning or at least interdependencies using data and text mining methods is well understood. Numerous approaches have been taken to select relevant information from often very large data sets. The discarding of items that are not relevant to a parameterized retrieval is usually based on an include or do not include decision imbedded in some kind of branch-and-bound algorithm, made to a varying extent sophisticated by the use of machine learning techniques. This paper addresses the discarding process as noise elimination within the context of well-established signal processing methods. It proposes an entropy-based approach using a value-weighted matrix for word relevance matching, where whole text is partitioned according to whether there is a direct relevance of word pairs to the declared meaning being sought, which is expressed as a set of parameters and the noise is considered as errors in the data stream. The resulting non-noisy data is depicted as a text meaning vector, where terms of direct relevance to the initial parameter values are stored.

Keywords—text mining; sentiment analysis; topic modeling; social media; entropy

I. INTRODUCTION

The approach to determining meaning in text that is proposed in this paper is early stage work in progress. It relates more generally to the research domain for what has become known as *social media* [1]. There are numerous perspectives on how social media is being used in popular culture and what effects it might have on personal, professional, corporate and political thinking and decision making. At the root of the issues around interpretation lays time-honoured linguistic phenomena that for computational processing have long dogged research in this domain. Some of these issues include ambiguity and dis-ambiguity, anaphoric reference (the association between pronouns where multiple nouns are present in the same sentence or paragraph), perplexity and redundancy. While not addressing these issues specifically in this paper they are nonetheless accounted for in the research work to which it relates. This paper in contrast, specifically deals with the issue of entropy and proposes a perspective on semantic processing of natural language (English) where this concept could be useful as an analytical instrument.

Entropy is a measure of unpredictability. English language text has low entropy and is therefore, fairly predictable. Even if we don't know exactly what is going to come next, we

can be fairly certain that, for example, there will be many more e's than z's, or that the combination 'qu' will be much more common than any other combination with a 'q' in it and the combination 'th' will be more common than any of them.

Entropy in information theory, as defined by Shannon [2] is a measure of the uncertainty associated with a random variable. In this context entropy is a quantification of the expected information value in a given text. His seminal work in this area demonstrated how we can measure the average information content that can be missed if one does not know the value of the random variable. So entropy in this context means we disambiguate the message in a text and by so doing receive the precise information necessary to understand it clearly.

Without broadening the discussion of entropy per se, because it has specific meaning in thermodynamics and several other contexts not related to text processing or meaning representation, it is noteworthy for analytical purposes to consider the expression

$$S = -k_b \sum_i P_i \ln P_i \quad (1)$$

that provides a logarithmic measure for statistical purposes when considering the density of states in a given data set. Where a number of discrete yet related variables occur, we can observe that entropy exists in the sub-set of elements that are in a non-relevant state according to the meaning parameters defined at the outset of processing. For the purposes of this paper, we consider this to be a 'noisy' state.

So, measuring uncertainty such that we can identify entropy, or noisy text in a message, means we first need to identify the random variable (X below) and then using a probability mass function across the complete set, observe the logarithmic progression, which will in turn enable a vector of relevant (meaningful) terms to be generated.

$$H(X) = - \sum_i p(x_i) \log p(x_i) \quad (2)$$

We can extend this to include conditional entropy (X and Y below) so that our model can be choice-based when certain conditions are met. This assists in the term classification process and increases the robustness of the selection of noisy and non-noisy words or phrases.

$$H(X|Y) = - \sum_i \sum_j p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \quad (3)$$

The entropy measure using Shannons approach [3] provides us with a value that is assigned to the realized information content contained in a message. This is the unpredictable value we have extracted using the random variable method and observed as a logarithmic progression. This is considerably richer (more meaningful) than what we can determine at first glance as the predictable or known part of the text that we regard as meaningful. Part of the disambiguation of the text is to eliminate redundancy, which can create confusion for the reader and promote uncertainty in any statistical analysis of the text structure where properties such as the number of adjacent pairs of words occur repetitively and therefore, invalidate word frequency and spatial mapping of the text. We can use a Markov process to assist here in eliminating this particular noise from the data stream.

In simple form the Markov expression in Equation 4 is a common way to define entropy for text where p_i is the probability of i .

$$H(S) = - \sum_i p_i \log p_i \quad (4)$$

This may be extended to include first and second-order Markov components:

$$H(S) = - \sum_i p_i \sum_j p_i(j) \sum_k p_{i,j}(k) \log p_{i,j}(k) \quad (5)$$

Where w_i is a state (as defined previously for certain preceding characters) and $p_i(j)$ is the probability of j given i as the previous word.

II. LATENT TOPIC OPINION MINING

In this paper, we address the problem of opinion analysis using a probabilistic approach for the underlying structure of different types of opinions or sentiments around a certain object. Probabilistic models such as *topic models* can be used for discovering the hidden or latent description or the topic of a group of opinions using a particular combination of words [4]. In topic modeling, a term-document matrix X is extracted from a set of opinions around a particular subject. This matrix describes the occurrences of terms in opinions and is composed the frequencies on each one of the phrases, so each element x_{ij} contains the frequency of the word w_i in the opinion o_j .

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

$$p(o) = \sum_z p(z) \prod_{n=1}^N (p(w_n|z)) \quad (6)$$

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.

$$p(o, w_n) = p(o) \sum_z p(z) p(w_n|z) p(z|o) \quad (7)$$

Each opinion is then represented as a list of mixture proportions representing its membership to any particular topic. Due to the unigram 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 specially 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) [5] 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 infinitely exchangeable topics, so the probability of a sequence of words and topics can be written as:

$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^N p(z_n|\theta) p(w_n|z_n) \right) d\theta \quad (8)$$

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

```

III. OPINION MINING AS A CASE OF TOPIC MODELING

In data mining, the problem of analyzing sentiment or emotion in text is called opinion mining. Using this technique, we would like to be able to classify a review or an opinion of any particular subject. From a decision maker’s point of view having a notion of the trend or the sentiment around a subject in millions of opinions can be detrimental [6].

In general, opinions can be made on products, services or events. One of the issues that arises when analyzing and classifying opinions is the natural subjectivity of the process. Any qualitative judgement of an object might not be completely classified as being ‘bad’ or ‘good’, or might not contain any useful information about the subject.

For example, when analyzing sentiments about the iPad in Twitter we can find the phrase:

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

Which can be classified as a good sentiment. However, phrases like:

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

The above phrase contains no valuable information that can be used to analyze the emotion of the users around the iPad. In such cases, we would like to be able to detect whether any piece of text contains no useful information so we can remove this when building a classifier system.

IV. A SIGNAL DENOISING ALGORITHM FOR OPINION SPAM DETECTION

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:

$$p(O_M) = \exp\left(-\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d}\right) \quad (9)$$

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
```

V. CONCLUSION

This paper sets out a framework for using a signal denoising algorithm that can be applied to large sets of continuous noisy data such as that coming from social media text. It proposes that an entropy-based approach be used with a mixture modeling technique for spam mining in opinion datasets. The approach appears to be robust mathematically, although there are questions yet to be addressed in terms of linking the elements of the approach into a unified methodology. The principal challenge here is to apply the method to large data sets in order to observe the elimination rate of errors and indeed, whether the elimination is accurate in terms of the residual message comprehension and meaning.

REFERENCES

- [1] J. Surowiecki, *The Wisdom of Crowds*. New York: Anchor Books, 2004.
- [2] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press, 1949.
- [3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [4] B. Liu, “Sentiment analysis: A multifaceted problem,” *IEEE Intelligent System*, vol. 25, no. 3, pp. 76–80, 2010.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res*, vol. 3, no. 3, pp. 993–1022, 2003.
- [6] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)*, Stanford, California, USA, Feb 2008.