An Information Theoretic Approach to Measuring Text Interestingness

Anonymous Author(s)
Affiliation
Address
email

Abstract

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

2 Our Approach

2.1 Information Diversity

As a starting point of our model, we assume that we can describe any word with a probability distribution over a fixed set of topics - e.g. learned from a topic model. Now, given a small set of words, we in effect have a set of topic distributions. Let us denote the set of words by $S = \{w_1, ..., w_k\}$, and the distribution of a word w by P_w. We now ask what is the diversity of the set of probability distributions $\mathcal{P}_S = \{P_{w_1}, ..., P_{w_k}\}$? To make this task somewhat more concrete, we make some assumptions of what we expect from the measure. First, if the set $\mathcal{P}_S$ only contains one element, then it is not diverse. Moreover, we have a certain fixed *prior* distribution P, which we can think of as the overall distribution of topics. If a distribution P_{w_i} is equal to P, then it does not carry any information, so it should not have any impact on the diversity of $\mathcal{P}_S$. In accordance with those assumptions, we will will introduce some basic notation. Let $D_{KL}(.\|.)$ denote the Kullback-Leibler Divergence.

Definition 1 *Given a distribution P_w, its* **importance** *with respect to a prior distribution P is defined as*

$$D_w = D_{KL}(P_w \| P).$$

Definition 2 *Given a set of distributions $\mathcal{P}_S = \{P_{w_1}, ..., P_{w_k}\}$ and a prior P, we define a mixture distribution P_S as*

$$P_S = \sum_{i=1}^{k} d_{w_i} P_{w_i},$$

where $d_{w_i} = \frac{D_{w_i}}{\sum D_{w_j}}$ are the normalized importances.

Essentially, P_S is the weighted average of the set $\mathcal{P}_S$, where the weights are chosen according to the importances. Now, we can define the diversity measure.

Definition 3 *We define the Jensen-Shannon Information Diversity of a set of distributions $\mathcal{P}_S$ with respect to prior P as*

$$D_S = \sum_{i=1}^{k} d_{w_i} D_{KL}(P_{w_i} \| P_S),$$

where d_{w_i} and P_S are as in the previous definition.

This definition is closely related to the *general Jensen-Shannon Divergence*, defined in [topsoe]. Another interesting theoretical property of Jensen-Shannon Information Diversity is that it can be interpreted as a generalization of Shannon entropy as a population diversity measure, however we will not go into this here any further.

2.2 Text Diversity

Applying this general model to the natural language requires some additional considerations. First, how do we obtain a topic ditribution for a given word, and what is the prior topic distribution. As the terminology suggests, we want to rely on a topic model. Specifically, suppose we have a set $\mathcal{T}$ of documents, sentences or paragraphs of any length, in which we want to find the most diverse examples. Additionally, we have a large set $\mathcal{D}$ of documents which has content *similar* to $\mathcal{T}$. We will train a Latent Dirichlet Allocation (LDA) topic model on $\mathcal{D}$, and use it to analyze $\mathcal{T}$. Of course, both sets could be identical, but $\mathcal{T}$ may, for example, be a set of sentences or short messages, which not suitable for training a good topic model on. In those cases, as we will see, the fact that $\mathcal{T}$ and $\mathcal{D}$ do not need to be identical will give our method a distinct advantage over competing approaches. The key component we will need from the topic model is the word-topic matrix M, where M_{ij} is the number of times i-th word was assigned the j-th topics. So, if we probabilistically normalize the i-th row of M, it will be the topic assignment distribution for the i-th word in our dictionary. This is a good candidate for a topic distribution to use in our model. Similarly, we obtain a prior topic distribution by computing the proportions of overall topic assignments. This corresponds to summing up matrix M along its columns, and then normalizing the resulting vector.

The proposed choice of topic distributions for words has some shortcomings. First, consider a word w which occurs only once (or very few times in the entire set $\mathcal{D}$. Then, its topic distribution will be very concentrated. Assuming that the prior distribution is close to uniform, that will give a very high importance to w, just because there is not enough datapoints in our set to determine its *true* topic distribution.

Furthermore, if we interpret a topic distribution as representing the *meaning* of a word, then having a fixed distribution for each word would not be able to reflect the fact that words often have different meanings in different contexts. In effect, all of the meanings would be combined together, and likely the most frequent one would dominate, distorting the results.

Finally, there is a fundamental problem with describing the meaning of words by their topic distributions. It forces us to make an implicit assumption that the topics form a basis that is, in certain sense, orthogonal. Consider, for example, a word that has a narrow usage, so that it is always assigned the same topic in our model. Then, its distribution will show no relation with any other topics, even those that may be closely related to it. This intuitively seems like a misrepresentation of the meaning of that word. This issue was also raised in [topsoe]. The solution proposed there incorporated topic similarities into a new diversity measure. We will present a different approach that also uses topic similarities.

2.3 Observer's Model

We will now present a model of a human Observer A that is being presented with the set $\mathcal{T}$ and is supposed to judge the *interestingness* of each piece of text from it. Additionally, we assume that A has gained their linguistic knowledge by observing samples from the set $\mathcal{D}$. What is, then, the general meaning (topic distribution) of a word w, from the Observer's viewpoint? A statistically adequate approach is to take a Bayesian model with a Dirichlet prior. We observe occurrences in set $\mathcal{D}$ of topic assignments for w to learn the posterior distribution. However, instead of using the fixed assignments produced by the topic model, we let the Observer choose their own assignment as

follows: given a pair (w, t), where t is the topic assignment given by the model for a specific occurrence of w, the observer selects a topic from some topic-similarity distribution S_t, conditional on t. An appropriate Dirichlet prior can be derived from this by looking at the overall topic assignments (not for a specific word). What would we get if we fed those to the Observer, letting A generate their own topic for each? Denoting by S the matrix with rows of topic-similarity distributions and by P the (row vector) topic distribution coming from the word-topic matrix, we find that the observer's prior will concentrate around the distribution described by the product $\widehat{P} = PS^T$. Using Bayes's rule to calculate the Observer's posterior probability, we get

$$\widehat{P}_w = \frac{\alpha PS^T + \mu_w P_w S^T}{\alpha + \mu_w},$$

where μ_w is the frequency of word w in D, and P_w is the topic assignment distribution obtained from the word-topic matrix, while α is the parameter that specifies the strength of the prior.

Next, let us analyze the Observer's behavior when reading a text segment from $\mathcal{T}$. We treat each piece of text as a bag of words, disregarding the order. Suppose, the set of words is $T = \{w_1, ..., w_k\}$. We have already established how the Observer understands each word separately. However, given a set of words, each one exists in the context of the others. We can describe that context using the mixture distribution from Definition 2. Denote $T_1 = \{w_2, ..., w_k\}$ as the set of all words in T except w_1. What is the appropriate topic distribution for w_1, given a context mixture distribution P_{T_1}? For this, we can look more closely at the LDA model we used to obtain the word-topic matrix. We can think of it as being generated by the following process: first drawing a topic form the prior topic distribution, then drawing a word from that topic's word distribution. It is natural to ask what would the matrix look like if we used P_{T_1} as the topic distribution instead of the prior, and what would be the corresponding topic distribution for word w_1.

Proposition 4 *Let $\widehat{P}, \widehat{P}_{w_1}, P_{T_1}$ be the topic prior, general topic distribution for w_1, and the context distribution, respectively. Then, the context dependent distribution defined as above, will be*

$$P_{w_1}^{T_1}(t) \propto \frac{\widehat{P}_{w_1}(t) P_{T_1}(t)}{\widehat{P}(t)}.$$

The danger with relying on $P_{w_1}^{T_1}$ is that if the distributions $\widehat{P}_{w_1}$ and P_{T_1} are mostly disjoint, than their product will be very small. In other words, we would need a long sampling process in generating the hypothetical word-topic count matrix to obtain a statistically significant estimation of the $P_{w_1}^{T_1}$. Once again, we turn to bayesian analysis: we let the Observer use $\widehat{P}_{w_1}$ as their maximum likelihood estimator in a Dirichlet prior, obtaining the following posterior solution:

$$\widehat{P}_{w_1}^{T_1}(t) \propto \beta \widehat{P}_{w_1}(t) + \frac{\widehat{P}_{w_1}(t) P_{T_1}(t)}{\widehat{P}(t)}.$$

3 Experiments

3.1 Data Sets

- **iPhone cases:**
- **NSF:**

3.2 Baselines

Diversity:

- **Shannon Entropy:**
- **Topic Diversity:** [**?**], [**?**]
- **Word frequency as distribution:**

Classification:

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

- **Bag of words (BOW):**
- **Latent Semantic Indexing (LSI):**
- **Recursive Auto Encoders (RAE):**

3.3 Results

4 Conclusions

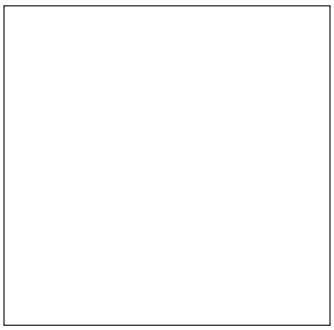

Figure 1: Sample figure caption.