Are These Documents Written from Different Perspectives? A Test of Different Perspectives Based On Statistical Distribution Divergence

Wei-Hao Lin
Carnegie Mellon University

Alexander Hauptmann
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/compsci

Published In
Abstract

In this paper we investigate how to automatically determine if two document collections are written from different perspectives. By perspectives we mean a point of view, for example, from the perspective of Democrats or Republicans. We propose a test of different perspectives based on distribution divergence between the statistical models of two collections. Experimental results show that the test can successfully distinguish document collections of different perspectives from other types of collections.

1 Introduction

Conflicts arise when two groups of people take very different perspectives on political, socio-economical, or cultural issues. For example, here are the answers that two presidential candidates, John Kerry and George Bush, gave during the third presidential debate in 2004 in response to a question on abortion:

(1) Kerry: What is an article of faith for me is not something that I can legislate on somebody who doesn’t share that article of faith. I believe that choice is a woman’s choice. It’s between a woman, God and her doctor. And that’s why I support that.

(2) Bush: I believe the ideal world is one in which every child is protected in law and welcomed to life. I understand there’s great differences on this issue of abortion, but I believe reasonable people can come together and put good law in place that will help reduce the number of abortions.

After reading the above transcripts some readers may conclude that one takes a “pro-choice” perspective while the other takes a “pro-life” perspective, the two dominant perspectives in the abortion controversy.

Perspectives, however, are not always manifested when two pieces of text together are put together. For example, the following two sentences are from Reuters newswire:

(3) Gold output in the northeast China province of Heilongjiang rose 22.7 pct in 1986 from 1985’s level, the New China News Agency said.

(4) Exco Chairman Richard Lacy told Reuters the acquisition was being made from Bank of New York Co Inc, which currently holds a 50.1_pct, and from RMJ partners who hold the remainder.

A reader would not from this pair of examples perceive as strongly contrasting perspectives as the Kerry-Bush answers. Instead, as the Reuters annotators did, one would label Example 3 as “gold” and Example 4 as “acquisition”, that is, as two topics instead of two perspectives.

Why does the contrast between Example 1 and Example 2 convey different perspectives, but the contrast between Example 3 and Example 4 result in different topics? How can we define the impalpable “different perspectives” anyway? The definition of “perspective” in the dictionary is “subjective evaluation of relative significance,” but can we have a computable definition to test the existence of different perspectives?

1The American Heritage Dictionary of the English Language, 4th ed. We are interested in identifying “ideological perspectives” (Verdonk, 2002), not first-person or second-person “perspective” in narrative.
The research question about the definition of different perspectives is not only scientifically intriguing, it also enables us to develop important natural language processing applications. Such a computational definition can be used to detect the emergence of contrasting perspectives. Media and political analysts regularly monitor broadcast news, magazines, newspapers, and blogs to see if there are public opinion splitting. The huge number of documents, however, make the task extremely daunting. Therefore an automated test of different perspectives will be very valuable to information analysts.

We first develop statistical models for the two document collections, $A$ and $B$, and then measure the degree of contrast by calculating the “distance” between $A$ and $B$. How document collections are statistically modeled and how distribution difference is estimated are described in Section 3. The document corpora are described in Section 4. In Section 5, we evaluate how effective the proposed test of difference perspectives based on statistical distribution. The experimental results show that the distribution divergence can successfully separate document collections of different perspectives from other kinds of collection pairs. We also investigate if the pattern of distribution difference is due to personal writing or speaking styles.

2 Related Work

There has been growing interest in subjectivity and sentiment analysis. There are studies on learning subjective language (Wiebe et al., 2004), identifying opinionated documents (Yu and Hatzivasiloglou, 2003) and sentences (Riloff et al., 2003; Riloff and Wiebe, 2003), and discriminating between positive and negative language (Turney and Littman, 2003; Pang et al., 2002; Dave et al., 2003; Nasukawa and Yi, 2003; Morinaga et al., 2002). There are also research work on automatically classifying movie or product reviews as positive or negative (Nasukawa and Yi, 2003; Mullen and Collier, 2004; Beineke et al., 2004; Pang and Lee, 2004; Hu and Liu, 2004).

Although we expect by its very nature much of the language used when expressing a perspective to be subjective and opinionated, the task of labeling a document or a sentence as subjective is orthogonal to the test of different perspectives. A subjectivity classifier may successfully identify all subjective sentences in the document collection pair $A$ and $B$, but knowing the number of subjective sentences in $A$ and $B$ does not necessarily tell us if they convey opposing perspectives. We utilize the subjectivity patterns automatically extracted from foreign news documents (Riloff and Wiebe, 2003), and find that the percentages of the subjective sentences in the bitterlemons corpus (see Section 4) are similar (65.6% in the Palestinian documents and 66.2% in the Israeli documents). The high but almost equivalent number of subjective sentences in two perspectives suggests that perspective is largely expressed in subjective language but subjectivity ratio is not enough to tell if two document collections are written from the same (Palestinian v.s. Palestinian) or different perspectives (Palestinian v.s. Israeli)².

3 Statistical Distribution Divergence

We take a model-based approach to measure to what degree, if any, two document collections are different. A document is represented as a point

²However, the close subjectivity ratio doesn’t mean that subjectivity can never help identify document collections of opposing perspectives. For example, the accuracy of the test of different perspectives may be improved by focusing on only subjective sentences.
in a \(V\)-dimensional space, where \(V\) is vocabulary size. Each coordinate is the frequency of a word in a document, i.e., term frequency. Although vector representation, commonly known as a bag of words, is oversimplified and ignores rich syntactic and semantic structures, more sophisticated representation requires more data to obtain reliable models. Practically, bag-of-word representation has been very effective in many tasks, including text categorization (Sebastiani, 2002) and information retrieval (Lewis, 1998).

We assume that a collection of \(N\) documents, \(y_1, y_2, \ldots, y_N\) are sampled from the following process,

\[
\theta \sim \text{Dirichlet}(\alpha) \quad y_i \sim \text{Multinomial}(n_i, \theta).
\]

We first sample a \(V\)-dimensional vector \(\theta\) from a Dirichlet prior distribution with a hyperparameter \(\alpha\), and then sample a document \(y_i\) repeatedly from a Multinomial distribution conditioned on the parameter \(\theta\), where \(n_i\) is the document length of the \(i\)th document in the collection and assumed to be known and fixed.

We are interested in comparing the parameter \(\theta\) after observing document collections \(\mathcal{A}\) and \(\mathcal{B}\):

\[
p(\theta|\mathcal{A}) = \frac{p(\mathcal{A}|\theta)p(\theta)}{p(\mathcal{A})} = \text{Dirichlet}(\theta|\alpha + \sum_{y_i \in \mathcal{A}} y_i).
\]

The posterior distribution \(p(\theta|\cdot)\) is a Dirichlet distribution since a Dirichlet distribution is a conjugate prior for a Multinomial distribution.

How should we measure the difference between two posterior distributions \(p(\theta|\mathcal{A})\) and \(p(\theta|\mathcal{B})\)? One common way to measure the difference between two distributions is Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), defined as follows,

\[
D(p(\theta|\mathcal{A})||p(\theta|\mathcal{B})) = \int p(\theta|\mathcal{A}) \log \frac{p(\theta|\mathcal{A})}{p(\theta|\mathcal{B})} d\theta.
\]

Directly calculating KL divergence according to (5) involves a difficult high-dimensional integral. As an alternative, we approximate KL divergence using Monte Carlo methods as follows:

1. Sample \(\theta_1, \theta_2, \ldots, \theta_M\) from \(\text{Dirichlet}(\theta|\alpha + \sum_{y_i \in \mathcal{A}} y_i)\).

2. Return \(\hat{D} = \frac{1}{M} \sum_{i=1}^{M} \log \frac{p(\theta_i|\mathcal{A})}{p(\theta_i|\mathcal{B})}\) as a Monte Carlo estimate of \(D(p(\theta|\mathcal{A})||p(\theta|\mathcal{B}))\).

Algorithms of sampling from Dirichlet distribution can be found in (Ripley, 1987). As \(M \rightarrow \infty\), the Monte Carlo estimate will converge to true KL divergence by the Law of Large Numbers.

4 Corpora

To evaluate how well KL divergence between posterior distributions can discern a document collection pair of different perspectives, we collect two corpora of documents that were written or spoken from different perspectives and one newswire corpus that covers various topics, as summarized in Table 1. No stemming algorithms is performed; no stopwords are removed.

| Corpus  | Subset       | \(|\mathcal{D}|\)  | \(|d|\) | \(V\)  |
|---------|--------------|--------------------|--------|--------|
| bitterlemons | Palestinian  | 290                | 748.7  | 10399  |
|         | Israeli      | 303                | 822.4  | 11668  |
|         | Pal. Editor  | 144                | 636.2  | 6294   |
|         | Pal. Guest   | 146                | 859.6  | 8661   |
|         | Isr. Editor  | 152                | 819.4  | 8512   |
|         | Isr. Guest   | 151                | 825.5  | 8812   |
| 2004 Presidential Debate | Kerry       | 178                | 124.7  | 2594   |
|         | Bush         | 176                | 107.8  | 2393   |
|         | 1st Kerry    | 33                 | 216.3  | 1274   |
|         | 1st Bush     | 41                 | 155.3  | 1195   |
|         | 2nd Kerry    | 73                 | 103.8  | 1472   |
|         | 2nd Bush     | 75                 | 89.0   | 1333   |
|         | 3rd Kerry    | 72                 | 104.0  | 1408   |
|         | 3rd Bush     | 60                 | 98.8   | 1281   |
| Reuters-21578 | ACQ         | 2448               | 124.7  | 14293  |
|         | CRUDE        | 634                | 214.7  | 9009   |
|         | EARN         | 3987               | 81.0   | 12430  |
|         | GRAIN        | 628                | 183.0  | 8236   |
|         | INTEREST     | 513                | 176.3  | 6056   |
|         | MONEY-FX     | 801                | 197.9  | 8162   |
|         | TRADE        | 551                | 255.3  | 8175   |

Table 1: The number of documents \(|\mathcal{D}|\), average document length \(|d|\), and vocabulary size \(V\) of the three corpora.

The first perspective corpus consists of articles published on the bitterlemons website\(^3\) from late 2001 to early 2005. The website is set up to “contribute to mutual understanding [between Palestinians and Israelis] through the open exchange of ideas”\(^4\). Every week an issue about the Israeli-Palestinian conflict is selected for discussion (e.g., “Disengagement: unilateral or coordinated?”), and a Palestinian editor and an Israeli editor each contribute one article addressing the

\(^3\)http://www.bitterlemons.org/
\(^4\)http://www.bitterlemons.org/about/about.html
issue. In addition, the Israeli and Palestinian editors interview a guest to express their views on the issue, resulting in a total of four articles in a weekly edition. The perspective from which each article is written is labeled as either Palestinian or Israeli by the editors.

The second perspective corpus consists of the transcripts of the three Bush-Kerry presidential debates in 2004. The transcripts are from the website of the Commission on Presidential Debates. Each spoken document is roughly an answer to a question or a rebuttal. The transcript are segmented by the speaker tags already in the transcripts. All words from moderators are discarded.

The topical corpus contains newswire from Reuters in 1987. Reuters-21578 is one of the most common testbeds for text categorization. Each document belongs to none, one, or more of the 135 categories (e.g., “Mergers” and “U.S. Dollars”). The number of documents in each category is not evenly distributed (median 9.0, mean 105.9). To estimate statistics reliably, we only consider categories with more than 500 documents, resulting in a total of seven categories (ACQ, CRUDE, EARN, GRAIN, INTEREST, MONEY-FX, and TRADE).

5 Experiments

A test of different perspectives is acute when it can draw distinctions between document collection pairs of different perspectives and document collection pairs of the same perspective and others. We thus evaluate the proposed test of different perspectives can thus be measured by how different perspectives can thus be measured by how the entropies of collection pairs of the same perspective and others.

To account for large variation in the number of words and vocabulary size across corpora, we normalize the total number of words in a document collection to be the same $K$, and consider only the top $C\%$ frequent words in the document collection pair. We vary the values of $K$ and $C$, and find that $K$ changes the absolute scale of KL divergence but does not change the rankings of four conditions. Rankings among four conditions is consistent when $C$ is small. We only report results of $K = 1000$, $C = 10$ in the paper due to space limit.

There are two kinds of variances in the estimation of divergence between two posterior distributions and should be carefully checked. The first kind of variance is due to Monte Carlo methods. We assess the Monte Carlo variance by calculating a 100$\alpha$ percent confidence interval as follows,

$$[\hat{D} - \Phi^{-1}(\frac{\alpha}{2})\frac{\hat{\sigma}}{\sqrt{M}}, \hat{D} + \Phi^{-1}(1 - \frac{\alpha}{2})\frac{\hat{\sigma}}{\sqrt{M}}]$$

where $\sigma^2$ is the sample variance of $\theta_1, \theta_2, \ldots, \theta_M$, and $\Phi(\cdot)^{-1}$ is the inverse of the standard normal cumulative density function. The second kind of variance is due to the intrinsic uncertainties of data generating processes. We assess the second kind of variance by collecting 1000 bootstrapped samples, that is, sampling with replacement, from each document collection pair.

5.1 Quality of Monte Carlo Estimates

The Monte Carlo estimates of the KL divergence from several document collection pair are listed in Table 2. A complete list of the results is omitted due to the space limit. We can see that the 95$\%$ confidence interval captures well the Monte Carlo estimates of KL divergence. Note that KL divergence is not symmetric. The KL divergence
Table 2: The Monte Carlo estimate $\hat{D}$ and 95% confidence interval (CI) of the Kullback-Leibler divergence of several document collection pairs $(A, B)$ with the number of Monte Carlo samples $M = 1000$.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>$\hat{D}$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACQ</td>
<td>ACQ</td>
<td>2.76</td>
<td>[2.62, 2.89]</td>
</tr>
<tr>
<td>Palestinian</td>
<td>Palestinian</td>
<td>3.00</td>
<td>[3.54, 3.85]</td>
</tr>
<tr>
<td>Palestinian</td>
<td>Israeli</td>
<td>27.11</td>
<td>[26.64, 27.58]</td>
</tr>
<tr>
<td>Israeli</td>
<td>Palestinian</td>
<td>28.44</td>
<td>[27.97, 28.91]</td>
</tr>
<tr>
<td>Kerry</td>
<td>Bush</td>
<td>58.93</td>
<td>[58.22, 59.64]</td>
</tr>
<tr>
<td>ACQ</td>
<td>EARN</td>
<td>615.75</td>
<td>[610.85, 620.65]</td>
</tr>
</tbody>
</table>

of the pair (Israeli, Palestinian) is not necessarily the same as (Palestinian, Israeli). KL divergence is greater than zero (Cover and Thomas, 1991) and equal to zero only when document collections $A$ and $B$ are exactly the same. Here (ACQ, ACQ) is close to but not exactly zero because they are different samples of documents in the ACQ category. Since the CIs of Monte Carlo estimates are reasonably tight, we assume them to be exact and ignore the errors from Monte Carlo methods.

5.2 Test of Different Perspectives

We now present the main result of the paper. We calculate the KL divergence between posterior distributions of document collection pairs in four conditions using Monte Carlo methods, and plot the results in Figure 1. The test of different perspectives based on statistical distribution divergence is shown to be very acute. The KL divergence of the document collection pairs in the DP condition fall mostly in the middle range, and is well separated from the high KL divergence of the pairs in DT condition and from the low KL divergence of the pairs in SP and ST conditions. Therefore, by simply calculating the KL divergence of a document collection pair, we can reliably predict that they are written from different perspectives if the value of KL divergence falls in the middle range, from different topics if the value is very large, from the same topic or perspective if the value is very small.

5.3 Personal Writing Styles or Perspectives?

One may suspect that the mid-range distribution divergence is attributed to personal speaking or writing styles and has nothing to do with different perspectives. The doubt is expected because half of the bitterlemons corpus are written by one Palestinian editor and one Israeli editor (see Table 1), and the debate transcripts come from only two candidates.

We test the hypothesis by computing the distribution divergence of the document collection pair (Israeli Guest, Palestinian Guest), that is, a Different Perspectives (DP) pair. There are more than 200 different authors in the Israeli Guest and Palestinian Guest collection. If the distribution divergence of the pair with diverse authors falls out of the middle range, it will support that mid-range divergence is due to writing styles. On the other hand, if the distribution divergence still fall in the middle range, we are more confident the effect is attributed to different perspectives. We compare the distribution divergence of the pair (Israeli Guest, Palestinian Guest) with others in Figure 2.

The results show that the distribution divergence of the (Israeli Guest, Palestinian Guest) pair, as other pairs in the DP condition, still falls in the middle range, and is well separated from SP and ST in the low range and DT in the high range. The decrease in KL divergence due to writing or speaking styles is noticeable, and the overall effect due to different perspectives is strong enough to make the test robust. We thus conclude that the test of different perspectives based on distribution divergence indeed captures different perspectives, not personal writing or speaking styles.

5.4 Origins of Differences

While the effectiveness of the test of different perspectives is demonstrated in Figure 1, one may
wonder why the distribution divergence of the
document collection pair with different perspec-
tives falls in the middle range and what causes the
large and small divergence of the document collec-
tion pairs with different topics (DT) and the same
topic (ST) or perspective (SP), respectively. In
other words where do the differences result from?

We answer the question by taking a closer look
at the causes of the distribution divergence in our
model. We compare the expected marginal dif-
fERENCE of $\theta$ between two posterior distributions
$p(\theta|\mathcal{A})$ and $p(\theta|\mathcal{B})$. The marginal distribution of
the $i$-th coordinate of $\theta$, that is, the $i$-th word in the
vocabulary, is a Beta distribution, and thus the ex-
pected value can be easily calculated. We plot the
$\Delta \theta = E[\theta_i|\mathcal{A}] - E[\theta_i|\mathcal{B}]$ against $E[\theta_i|\mathcal{A}]$ for each
condition in Figure 3.

How $\Delta \theta$ is deviated from zero partially explains
different patterns of distribution divergence in Fig-
ure 1. In Figure 3d we see that the $\Delta \theta$ increases
as $\theta$ increases, and the deviance from zero is much
greater than those in the Same Perspective (Fig-
ure 3b) and Same Topic (Figure 3a) conditions.
The large $\Delta \theta$ not only accounts for large distribu-
tion divergence of the document pairs in DT con-
ditions, but also shows that words in different top-
ics that is frequent in one topic are less likely to be
frequent in the other topic. At the other extreme,
document collection pairs of the Same Perspective
(SP) or Same Topic (ST) show very little difference in $\theta$, which matches our intuition that docu-
ments of the same perspective or the same topic
use the same vocabulary in a very similar way.

The manner in which $\Delta \theta$ is varied with the
value of $\theta$ in the Different Perspective (DP) con-
dition is very unique. The $\Delta \theta$ in Figure 3c is not
as small as those in the SP and ST conditions,
but at the same time not as large as those in DT
conditions, resulting in mid-range distribution di-
vergence in Figure 1. Why do document collec-
tions of different perspectives distribute this way?
Partly because articles from different perspectives
focus on the closely related issues (the Palestinian-
Israeli conflict in the bitterlemons corpus, or the
political and economical issues in the debate cor-
pus), the authors of different perspectives write or
speak in a similar vocabulary, but with emphasis
on different words.

6 Conclusions

In this paper we develop a computational test of
different perspectives based on statistical distribu-
tion divergence between the statistical models
of document collections. We show that the pro-
Figure 3: The $\Delta \theta$ vs. $\theta$ plots of the typical document collection pairs in four conditions. The horizontal line is $\Delta \theta = 0$.

Proposed test can successfully separate document collections of different perspectives from other types of document collection pairs. The distribution divergence falling in the middle range cannot simply be attributed to personal writing or speaking styles. From the plot of multinomial parameter difference we offer insights into where the different patterns of distribution divergence come from.

Although we validate the test of different perspectives by comparing the DP condition with DT, SP, and ST conditions, the comparisons are by no means exhaustive, and the distribution divergence of some document collection pairs may also fall in the middle range. We plan to investigate more types of document collections pairs, e.g., the document collections from different text genres (Kessler et al., 1997).

**Acknowledgment**

We would like thank the anonymous reviewers for useful comments and suggestions. This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number NBCHC040037.
References


