AUTOMATED BLOG CLASSIFICATION: CHALLENGES AND PITFALLS
AUTHORS

- Hong Qu
- Andrea La Pietra
- Sarah Poon
• Blogs are difficult to categorize by **humans and machines alike**, because they are written in a capricious style

• Number of blogs is growing at an exponential rate

• In 2004, the Pew Institute found that 2-7% of Internet users have a blog and 11% read blogs

• Technorati’s web crawlers indicate that there are about 12,000 new weblogs created each day (a new weblog created in every 7.4 seconds)

• Manually maintain blog directories became more and more impossible

• Given the popularity of blogs, it would be useful if we could devise a content classification system to automatically generate a directory of blogs

• It is difficult to group blogs into categories because of the freestyle nature of the discourse, bloggers write whatever is on their mind, sometimes inventing new vocabulary and grammar

• Some bloggers intentionally deviate from rules of language and decorum to create a spectacle intending to attract a larger audience.
OVERVIEW OF EXPERIMENT

- Investigated the efficacy of using machine learning to categorize blogs
- Tried to classifying blogs using pure statistical measures such as TF-IDF
- Experimented with giving more weight to linguistic features such as the title of the blog posts and the anchor text from incoming links
- Ineffective, because blogs do not fit neatly in mutually exclusive Categories (a particular blog can fall into multiple Categories)
- Design a text classification experiment to categorize 120 blogs into 4 topics
  - personal diary
  - News
  - Political
  - sports.
- The baseline feature is unigrams weighed by TF-IDF, which yielded 84% accuracy
- Analyzed the corpus, features, and result data

Findings: Blog taxonomies need to support polyhierarchy, a given blog may be correctly classified under more than one category
PREVIOUS WORK

• Krishnamurthy proposes a classification system along **two dimensions**: personal vs. topical, and individual vs. community
• Hobbyist and experts write topical blogs
• Personal blogs, on the other hand, are written as a personal newsletter for the benefit of family, friends, and random strangers
• Topical blogs have clearly delineated topics, because the audience expects the blogger to be on topic
• Personal bloggers are not confined to one topic, they tend to meander across a range of topics, which has implications for content classification.
METHODOLOGY-PROCESSING THE CORPUS

• First, limited the scope of topics to four topic categories (personal diary, news, politics, and sports)
• Manually harvested and classified 30 blogs for each categories
• Parsed the RSS feeds from these blogs
• Used an open source RSS tool called Magpie to extract the title and body text from the blogs
• Finally, used NLTK and Weka to prepare the corpus and process the text
• Measured \textit{tf-idf} weigh of \textit{single word tokens} (unigrams) in the corpus
• Removed stop words
• Used the \textit{Naïve Bayes Multinomial classification} algorithm because it was the fastest and most accurate Weka classifier !!! Other Weka classifiers yielded significantly lower accuracy.
# EVALUATIONS

<table>
<thead>
<tr>
<th>Linguistic Feature</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams (baseline)</td>
<td>84%</td>
</tr>
<tr>
<td>Title + 1st sentence</td>
<td>76%</td>
</tr>
<tr>
<td>Anchor text</td>
<td>80%</td>
</tr>
<tr>
<td>Title + 1\textsuperscript{st} sentence and Anchor text</td>
<td>73%</td>
</tr>
</tbody>
</table>
EVALUATIONS

• closer examination of results form the training data set and the testing data set, a peculiar result is noticed-
  almost all the blogs were correctly identified except for political and news blogs

• Results indicates that, the category news blogs is very difficult to pin down. News blogs often talk about politics. Even human judgment would have trouble determining whether a blog that talks mostly about politics is a political or a news blog.
## Evaluations

### Training Dataset Results

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>&lt;- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>a = personal</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>b = news</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>c = political</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>17</td>
<td></td>
<td>d = sports</td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---
# EVALUATIONS

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>a = personal</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>b = news</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>c = political</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>d = sports</td>
</tr>
</tbody>
</table>
INSIGHTS

• Source of the erroneous classifications stem from a flawed taxonomy
• The topic boundary between news and politics is blurry. Yet the blogosphere is full of blogs that address multiple topics.
• Consequently, the first step in building an automated blog classification system—taxonomy design—is a pitfall because some blogs belong in multiple categories
NEXT STEPS

• Should have used a faceted classification that allows for polyhierarchy
• One approach for designing a polyhierarchical blog classification system could be to
  • divide a blog into individual posts.
  • the algorithm would classify individual blog posts
  • then applying a percentage threshold to determine which facet(s) to assign the blog.
• The benefit of this approach is that a blog post is likely to be limited to one topic
• The drawback is that a single post is a small document and is therefore more difficult to classify
THANK YOU