WEIR: Critique

Kai Hayashi, Cary Lee, & Daniel Myers

J. Heide and N. Talpes determine the level of anger in a blog post via a linear regression applied to the word vector of the text body of the blog post. This regression is run on a collection hand labelled blogs posts consisting of 230 posts. These articles are rated on a scale of 1-10.

Critique of Method:

Use of Linear Regression:

Utilizing a linear regression model on term frequency to compute sentiment assumes that there is a linear relationship between the term frequency and the articles anger. However, the number of times that a word occurs in a blog post may not be linearly related to the actual sentiment.

Take, for example, Malcom X's speech "The Bomb or the Bullet". This speech uses several over-arcing motifs such as "ballet", and "bullets", but These words are not necessarily correlated to anger: he could, for example just be talking about dangerous ballet boxes in the south.

It is dangerous to take words out of context and base the linear regression solely off of these.

In addition, you limit the number of words utilized to 100 with a word frequency of 10. How are these 100 words chosen? If they are chosen by frequency then a substantial number of important words maybe eliminated.

Data Sources:

From a collection of 2000 blog posts, the authors take a hand identified subset of these posts to run their model on. To ensure that this data is good, several things must be ensured. First, the distribution of these 230 blog posts must be even (or close to) across blog genres. From your example words of what was indicative of anger, it seems like the blogs retrieved mainly focus on technology. However, there are many political and personal blogs that would be very angry that would never use words like 'kindle'. We are curious to see what the distribution of your blogs sources are. This may play into the reason for why your average anger level is so low.

Hand-labeling blog posts may introduce some bias depending on how the authors accomplished this task. If both of the authors gave all 230 blog posts a rating between 1-10 and then averaged then the issues is mute. However, there may be some personal bias if the labels were split between the two.

Suggestions:

Improvement of model:

To help improve the contextual issues we suggest that you utilize a 3-gram tokenizer in the string to word vector filter. This would preserve some of the context and maybe improve your results. However, it will increase the number of attributes that you use by a factor of 3.

Pay careful attention to the words filtered out by the string to word vector requirements. Some of them you may actually want to keep, requiring a higher limit on the word count.

It also seems like there maybe a nonlinear relationship between word frequency count and the sentiment of the posts, however it is difficult to tell from your results.

Improvement of Data:

From your average anger level, it is clear that you need more angry blogs. However, you do not want this effecting your test set. From the words that are indicative of anger it seems like you need a much wider range of blog topics. Perhaps mechanical turk may be a good way of assigning labels to these blogs, but pay attention to the average rating that is assigned to the blogs posts.

Stop words may also improve your performance.