Well, some find it funny: Exploring alternative methods of combining Humicroedit annotator scores

Dorijan Cirkveni

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
autor1@xxx.hr, {autor2, autor3}@zz.com

Abstract

Humor is an inherently subjective matter, as can be seen in the Humicroedit dataset's annotator grades which diverge greatly for higher averages, which this paper suggests might require an annotation analysis method different than those traditionally used on high-agreement datasets. In order to achieve a more evenly distributed labeling system, this paper tests a variety of annotator weights and a variety of methods to determine the ideal set of weights, and then compares performance on dataset with the regular mean.

1. Introduction

Artificial intelligence has been making major strides recently both in terms of development and practical usage, especially with the advent of LLMs such as ChatGPT as well as stable diffusion models such as Midjourney. However, certain tasks still remain outside its grasp.

One such task is the detection, analysis, and generation of humor, a task that continues to elude the methods used. And unlike humor recognition (Khodak et al., 2017; Davidov et al., 2010; Barbieri and Saggion, 2014; Reyes et al., 2012; Cattle and Ma, 2018; Bertero and Fung, 2016; Yang et al., 2015), humor generation using artificial intelligence has proven especially tasking. One obstacle in this line of research is the scarcity of public datasets, and even greater scarcity - if not outright absence - of topic-appropriate public datasets.

One large problem in this field of study is bias, because humor is as subjective as it is complex which makes it significantly challenging to tackle from a natural language analysis/synthesis standpoint.

There are two common approaches to this problem in terms of reducing domain size and therefore annotator bias, one being to focus on a specific domain (TODO: insert citations) and the other being to focus on a specific research topic.

The underlying paper of this paper, *"President Vows to Cut Taxes Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines* (Hossain et al., 2019), provides one dataset that is an example of the latter. A lot of care was put into eliminating annotator bias. However, as this paper suggests, these efforts might prove futile as humor is an inherently subjective activity unsuitable for such approaches.

In particular, this paper suggests alternative approaches to arithmetic mean when combining annotator grades, such as only considering top 2 grades or assigning custom weights to annotator scores.

2. Related Work

Alongside Humicroedit, there are other works such as A Large Self-Annotated Corpus for Sarcasm (Khodak et al., 2017), with a corpus that makes Humicroedit pale in comparison (1.3 million sarcastic statements - 10 times more than any previous dataset - and many times more instances of non-sarcastic statements), Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon, and Modelling Sarcasm in Twitter, a Novel Approach. However, this paper focuses entirely on Humicroedit and its attempt to curb annotator bias.

3. Humicroedit annotation methods

The Humicroedit dataset includes a list of 5, 10, or 15 annotator scores, with a vast majority only having 5 scores. However, a more manageable featureset was provided alongside it, featuring 5 annotator scores for every entry. These scores are integers from 0 to 3, and they are sorted in ascending order.

Unfortunately, these annotator sets do not appear to connect to individual annotators, making it difficult to model annotator profiles.

3.1. Annotation data properties

As mentioned in 1., Humicroedit was designed with an intention of minimizing annotator bias.. However, as noted in the paper, despite the writers' effort of carefully qualifying annotators, their perception of headlines in regards of humor present was influenced by their knowledge, preferences, bias and stance towards information presented in the headlines.

This bias is best seen in the annotator vote distribution, as lower mean scores are closer to unanimous than higher mean scores (as well as more common) as seen in Figure 1.

Annotator bias may also be presenting itself in topic frequency, as seen by one topic that appears in dataset entries more often than any other, as seen in Figure 2 at 39 percent of the total dataset size.

4. Distribution measurement methods

In order to effectively choose our weights, visual confirmation will not be a sufficient measure. In this paper we shall use 4 different methods of determining how even is the label distribution: the Chi-Square uniformity test, the

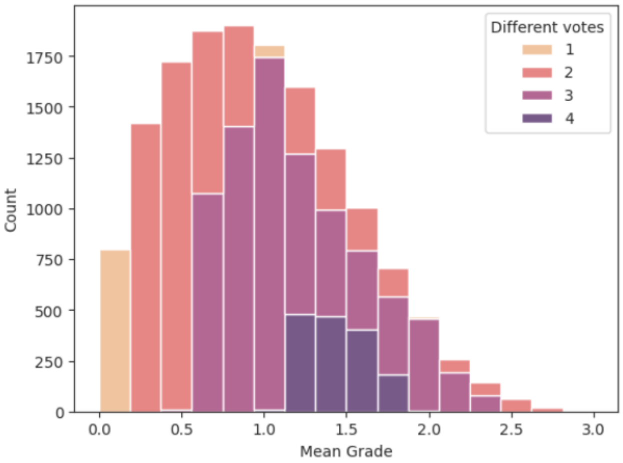

Figure 1: Distribution of votes for entire dataset.

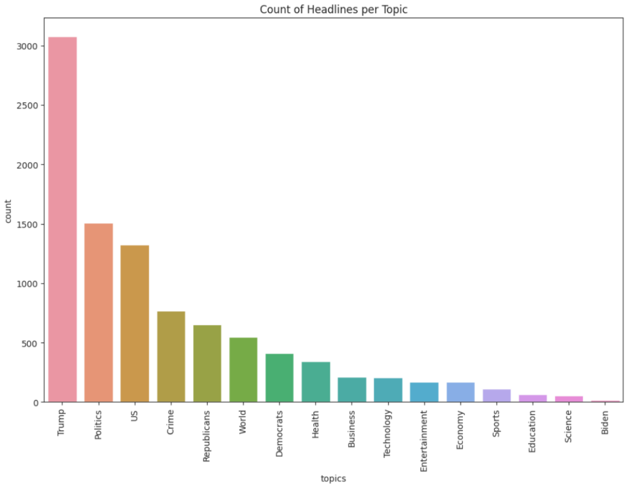

Figure 2: As we can see on this chart, a disproportionate amount of headline edits has been written on the topic of Donald J. Trump, a notable and controversial media personality and then-incumbent (45th) president of the United States of America.

Cramér-von Mises uniformity test, and a custom uniformity test developed specifically for this purpose.

4.1. Custom uniformity test

The custom uniformity test is based on the method we used to display unevenly wide histogram bars:

1. n is the number of unique values, and the number of bars. x_i is value of unique value i, and y_i is its frequency.

2. Set up limits between unique values to their arithmetic mean,

$$lim_i = (x_i + x_i + 1)/2, i > 0, i < n \tag{1}$$

3. Set up beginning and end limits in a way that first and last bars are centered on their values

$$lim_0 = x_0 * 2 - lim_1 \tag{2}$$

Table 1: Using methods on standard mean

Method	Result
Custom Uniformity	4.350
Chi Square Uniformity	5114.936
Cramér-von Mises Uniformity	1056.477

$$lim_n = x_n * 2 - lim_n - 1 \tag{3}$$

4. Calculate densities of data for each bar:

$$d_i = y_i/(lim_i + 1 - lim_i) i >= 0, i < n \tag{4}$$

5. Calculate reference density (for perfect set):

$$d_{total} = y/(lim_n - lim_0) \tag{5}$$

6. Apply weighted square mean error:

$$sme = \sum_{i=1}^{n} (lim_i + 1 - lim_i) * (d_i - d_{total})^2 \tag{6}$$

7. Divide result by reference density square:

$$sme_{adjusted} = sme/(lim_i + 1 - lim_i) * d_{total}^2 \tag{7}$$

4.2. Unweighted arithmetic mean (reference values)

First, we will determine the base distribution values by testing them on the provided labels, and the results of this are visible in Table 1.

These values will be useful once we compare the alternative methods to them.

5. Alternatives to simple aritmethic mean

5.1. Masking some of the annotator scores

Table 2: Custom uniformity

Data	Score
[0, 0, 0, 0, 1]	0.08799030892017554
[0, 0, 0, 1, 1]	0.33966016048738895
[0, 0, 1, 1, 1]	1.0731863586989547
[0, 1, 1, 1, 1]	2.3659776653569025
[1, 1, 1, 1, 1]	4.35032891060602

Table 3: Chi-Square uniformity

Data	Score
[0, 0, 0, 1, 0]	529.4772757387041
[0, 1, 0, 0, 1]	79.14734331522459
[0, 0, 1, 1, 1]	268.4623029018153
[0, 1, 1, 1, 1]	1218.7307539419637
[1, 1, 1, 1, 1]	5114.936593122639

First, we are going to use simple masking by only using certain annotator scores. Score masks are denoted as

Table 4: Cramer-Von Mises uniformity

Data	Score
$[0, 0, 1, 0, 0]$	1443.0052556034009
$[1, 0, 1, 0, 0]$	461.4334768687203
$[1, 1, 0, 1, 0]$	238.48263957896285
$[1, 1, 1, 1, 0]$	280.72257922049243
$[1, 1, 1, 1, 1]$	1056.4771337087336

follows: $[0, 1, 0, 1, 0]$, 0 and 1 used to mark ignored and considered scores respectively.

In this test, 31 ($2^5 - 1$) possible combinations are tested, sorted by result, divided into groups by number of scores used, and best result of each group is displayed, as lower number of annotator scores used may lead to less reliable results.

In Table 2, we can see the custom method produced more/less expected results, which is to say the highest annotator scores proved to provide the most uniform results.

However, Tables 3 and 4 tell a different story as not only do their scores give better scores for multiple annotators,

6. Future work

Unfortunately, due to time limitations brought upon by unfortunate circumstances and poor planning, we were unable to completely explore this topic. Future work in this line of reseach may involve further optimization of annotator grade weights towards providing more uniform results.

However, we recommend

7. Conclusion

8. Referencing literature

Unfortunately, due to compiling difficulties, we are unable to

9. Conclusion

Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

Acknowledgements

I would like to thank Gabriel Marinković for helping me try to make this paper.