Analysis of parent comments clusters with an anomalous sarcastic response

Marco Del Treppo - 961316

1 Introduction

The Sarcasm on Reddit dataset provides comments posted on Reddit labeled as sarcastic (1) or not sarcastic (0). The project goal is, given only the parent comment in a specific subreddit, identify parent comment clusters that receive an amount of sarcastic comments that deviate from the subreddit average. The project then focuses on document clustering, topic detection, and keyword extraction.

References

- Dimo Angelov, Top2Vec: Distributed Representations of Topics, [link]
- Grootendorst Maarten, BERTopic, [link]
- Grootendorst Maarten, Keyword Extraction with BERT, [link]
- L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering In: Journal of Open Source Software, The Open Journal, volume 2, number 11. 2017, [link]
- McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, [link]
- Universal Sentence Encoder, [link]

2 Research question and methodology

The research question is whether it is possible to identify parent comment clusters in which the mean of sarcastic responses exceeds the subreddit mean. The methodology is based on the references cited. In order to work with textual data, parent comments are transformed into vectors using the Universal Sentence Encoder. Then with UMAP the dimensionality of the data is reduced from 512 dimensions to just 5 dimensions. The reason for the reduction is to be able to use HDBSCAN as a clustering algorithm. Once the clusters are identified, the intra-cluster mean of the label variable is calculated to identify the clusters with the most sarcastic responses and consequently the topics that attract the most sarcastic responses.

3 Experimental results

The Sarcasm on Reddit dataset provides about one million scraped comments on Reddit.com. The comments were categorized, not by me, as sarcastic or not by leveraging the presence of the \s (sarcasm) tag. For analysis, the dataset was filtered on the "pcmasterrace" subreddit. The description of which provided by the creators themselves is as follows:

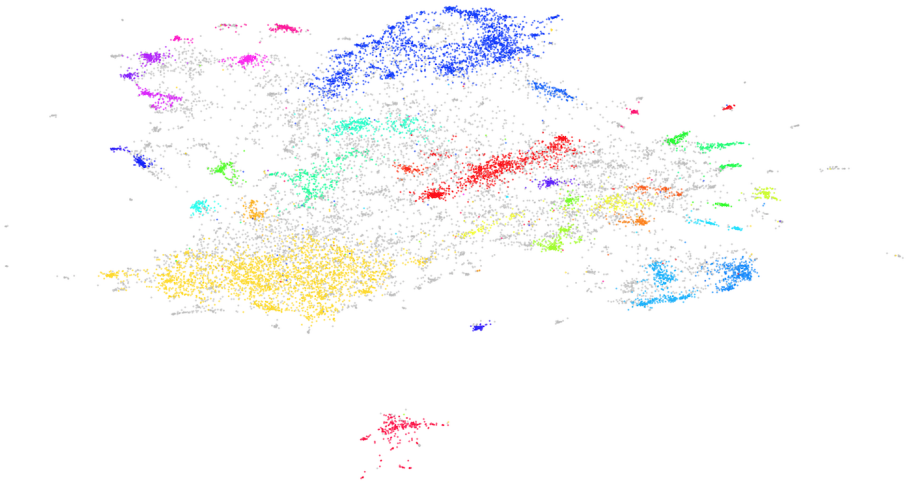
Welcome to the official subreddit of the PC Master Race. In this community, we celebrate and promote our favorite gaming and working platform. Ascend to a level that respects your eyes, your wallet, your mind, and your heart. Ascend to... the PCMR!

The dataset after being filtered and after having eliminated the columns not necessary for the analysis results to have 18999 rows and 2 columns. The columns are the parent comments (the comments that received the sarcastic response) and a categorical variable that indicates with 1 the presence of a sarcastic response and with 0 the absence.

Universal Sentence Encoder (USE)
As reported by the official page: "The Universal Sentence Encoder encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering and other natural language tasks. The model is trained and optimized for greater-than-word length text, such as sentences, phrases or short paragraphs. It is trained on a variety of data sources and a variety of tasks with the aim of dynamically accommodating a wide variety of natural language understanding tasks. The input is variable length English text and the output is a 512 dimensional vector."

Dimensionality Reduction and Clustering
USE returns vectors in 512 dimensions, UMAP, a fairly flexible non-linear dimension reduction algorithm, returns a reduced version in 5 dimensions of the

same vectors. HDBSCAN use unsupervised learning to find clusters, or dense regions, of a dataset. The following is a two-dimensional representation of the clusters and their parent comments. As is evident most of the clusters are colored gray. HDBSCAN does not force points into a cluster like other algorithms e.g. K-Mean. The gray points correspond to the group -1 or all points that have not been clustered. Clusters have variable size, but minimum size of 50 values. This threshold was defined in advance as a parameter of HDBSCAN in order not to have hundreds of small clusters.



Keyword Extraction Assuming that the extracted clusters can be a good proxy for the topics of the various parent comments, with the approach used so far we do not have a consistent technique to be able to define what is actually talked about in our clusters. One possibility is to compute the intra-cluster TF-IDF and use the most important words as headings.

The choice instead follows the nature of USE and sentence embeddings in general. Given the nature of vectors, it is possible to calculate the distance of two numerical vectors to give an assessment of similarity.

All tri-grams are then extracted from the comments and subsequently vectorized again using USE. Keeping in mind that we had a vector for each parent comment and we have the reference cluster available we will take the average intra-cluster vector as the reference vector. Calculating the distance between the average intra-cluster vectors and the vectors of the tri-grams we get the 5 tri-grams most similar to the topic of our cluster.

So we now have the various clusters, their size, and 5 tri-grams that give us

context in terms of the content of the various clusters. We finally calculate the average of the label for the various clusters thus determining if there are topics that attract more sarcastic comments. The extended results of the analysis can be extracted from the jupyter notebook.

Below are the mean values and related keywords of clusters with a mean greater than 0.6 or less than 0.4. Keeping in mind that the average of the subreddit is about 0.57 I think the results are at least interesting.

	Topic	Size	Keywords	Label
15	14	380	[hey 60 fps, fps confirmed 30fps, fps play 60fps, atleast getting 60fps, wait 60 fps]	0.789474
7	6	171	[mac laptops suck, macbooks make pc, macs just expensive, macs better pcs, macbook hate mac]	0.713450
24	23	59	[preordering stop complaining, whats point preordering, wont preorder just, fuck pre order, fuck preorder game]	0.694915
18	17	103	[yeah going upgrade, just little upgrade, got little upgrade, got nice upgrade, finally got upgrade]	0.689320
23	22	117	[better modded skyrim, modded skyrim just, heavily modding skyrim, modding skyrim amazing, modding skyrim feels]	0.683761
35	34	1004	[filthy console peasant, peasant thinks consoles, peasants think pc, peasant thinks pc, pc masterrace peasant]	0.680279
4	3	178	[g46 ram 12gb, just got ram, 6700k 16gb ram, 5820k 16gb ram, 12gb ram fucking]	0.674157
28	27	117	[ubisoft doesn want, does sorry ubisoft, admit ubisoft game, ubisoft likes game, really want ubisoft]	0.649573
9	8	77	[doesn use linux, linux use windows, linux distros people, just install linux, linux distros honestly]	0.649351
1	0	62	[vive putting psvr, oculus rift cv1, oculus 600 vr, vive vs oculus, oculus vr thing]	0.645161
20	19	309	[came pcmr brother, brothers glorious pc, brothers new pc, brothers today ascend, brothers today ascended]	0.605178
16	15	411	[wait 4k 144hz, 1440p 144hz 4k, 144hz wait 4k, 240hz 1080p monitor, 4k 60hz 1440p]	0.603406
12	11	113	[mechanical keyboard im, mechanical keyboard keyboard, wtf mechanical keyboards, keyboards really want, got mechanical keyboard]	0.495575
8	7	174	[just windows 10, windows 10 bad, windows 10 extremely, dreaded windows 10, windows 10 fuck]	0.494253
6	5	173	[3tb hardrive ssd, ssd 1tb hard, hdd better ssd, 500gb ssd just, hardrive ssd sadly]	0.491329
25	24	118	[internet shit isp, fucking data caps, glorious internet speeds, instantly internet speed, megabytes internet speed]	0.483051
2	1	342	[random game giveaway, giveaway steam keys, steam games giveaway, giveaway steam game, steam game giveaway]	0.219298

4 Concluding remarks

It is important to note that the results reported in the appendix are highly dependent on certain parameters chosen during the analysis that can be modified on the jupyter notebook. In particular they depend on ***n neighbors*** and ***n components*** parameters chosen in the dimensionality reduction. Subsequently from ***min cluster size*** the only parameter used in clustering with HDBSCAN.

From the analysis, it is evident that there are some topics, even well-defined ones, whose average sarcastic responses are far from the mean value. The critical point is that these clusters are generally very small and the larger clusters do not deviate from the mean value of the subreddit.

Another obvious flaw is the presence of a mega cluster of outliers -1 that accounts for nearly half of the observations. A possible speculation is that these parent comments are overly general comments, typical of forums, but to which

it is not possible to assign a topic that makes literal sense.

Solving the mega cluster problem and experimenting with new topic modeling solutions to this dataset are ideas for possible future work. Diversifying the extracted Keywords is also a possible value-add that can be useful to better define clusters.