Info 159 AP1

[Link to dataset](https://www.kaggle.com/datasets/usamabuttar/dad-jokes)

[Other dataset](https://www.kaggle.com/datasets/aryashah2k/dad-a-base-of-jokes)

**Description**

Dad jokes are simple, short, one- or two-line jokes that are commonly associated with cheesiness. Our group is curious about annotating these jokes and seeing how a model will perform in classifying the quality of these jokes compared to our personal opinions. To do this, we will be annotating dad jokes and categorizing them into the following: funny, lame, and cringy – based on how we subjectively view the jokes. For example, if a joke elicits a laugh or smile out of a reader, that is categorized as "funny". If a response is more lukewarm or not responsive at all (i.e. the reader understood the joke but does not find it funny), that corresponding joke would be categorized as "lame". If a negative response is elicited (such as a frown, or a reaction more responsive than a lukewarm response yet is not positive), that joke would be "cringy". Later in AP3, we will find commonalities between all the jokes we deem funny, lame, and cringe, and build our classifier on those commonalities. This is a task that requires human judgment because it is subjective based on the reader’s experience while reading the joke. Fluency in the language (here, English) is also required because many of these jokes require an understanding of phonetic puns, multiple meanings, and/or nuanced knowledge of particular words' sentiments. Our [data](https://www.kaggle.com/datasets/usamabuttar/dad-jokes) is sourced from Kaggle under a [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/) license, so the data is neither private nor under copyright. It contains ~7000 data points (which contain index and dad joke) and have 200 new entries added each day. The owner of the dataset did not classify how they collected the data. After cleaning our data of repeated jokes, we will pull 500 randomly selected jokes as our sample and classify them to our three labels.

**Datasheets**

1. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?
   1. No
2. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
   1. No
3. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
   1. No; dataset only contains index and dad jokes
4. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
   1. This dataset is a compilation of around ~7000 dad jokes, with around 200 added to it daily. For our purposes, we will only take 500 data points from the overall data set. In short, the dataset is not a sample from a larger set and rather is an actively growing dataset
   2. Randomly sample 500 data points
5. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
   1. Some of these points are either identical or extremely similar, so we will have to clean our data first and take out some data points before random sampling our 500 data points.
   2. We might also do some other methods of EDA/data cleaning on it, but we will also save the raw data for reference.
6. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
   1. Because we are taking a random sample of 500 rows from the larger dataset, collecting the data again could yield different results because different rows may be sampled randomly. Besides this, there is nothing that would impact future uses.
7. Are there tasks for which the dataset should not be used?
   1. These jokes are fabricated, so they shouldn’t be used to analyze formal language.
      1. Because many of these jokes may be puns, hidden meanings may be placed onto the subject of the joke, which may not actually reflect the true word-for-word meaning of the joke.
8. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
   1. No, this is a public dataset from Kaggle with a [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/) license.

**Example Document to Annotate**

A steak pun is a rare medium well done.