

How to Disagree-Ad Hominem
for
Knowledge Extraction and Information Retrieval

by

Marticola No: 942091

Amanpreet singh

Università degli Studi di Milano

TABLE OF CONTENTS

Abstract

1	How to disagree-Ad Hominem	1
1.1	Introduction	1
1.2	Research question and methodology	2
1.3	Experimental results	3
1.4	Concluding remarks	6

Chapter 1

How to disagree-Ad Hominem

1.1 Introduction

The project aims at automatically analyze corpuses in order to classify the counterarguments according to the Graham classification scheme. This project will focus on classifying Ad Hominem.

Ad Hominem

Ad Hominem is when, instead of addressing someone's argument or position, one irrelevantly attack the person or some aspect of the person who is making the argument. The fallacious attack can also be direct to membership in a group or institution. It is different from simply Name Calling as, in Ad Hominem one is attacking someone on the basis of their views, or what they said. Statement intended as a counter-argument is ad Hominem. If it's not an argument, it's not an ad Hominem argument.

Examples

"This does not logically follow. You evidently know nothing about logic."

"The structure of your argument is weak, you're saying idiotic things, therefore you're an idiot."

"We cannot approve of this recycling idea. It was thought of by a bunch of hippie communist weirdos."

1.1.1 Techinques and Methods Used

K-Means Clustering

K means clustering is Centroid based clustering. K-means clustering calculates the centroids and shuffles until it finds optimal centroid. It assumes that the number of clusters are already known (K). K-means follow Expectation Maximization approach to solve the problem. Expectation step is used for grouping the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster. The data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

NGRAMS

N-grams are simply all combinations of adjacent words or letters of length n that one can find in text. As an example, the "hello, world!" text contains the following word-level bigrams: "hello, world", "world, hello". The basic point of n-grams is that they capture the language structure from the statistical point of view, like what letter or word is likely to follow the given one. The longer the n-gram (the higher the n), the more context you have to work with. Optimum length really depends on the application. If n-grams are too short, it may fail to capture important differences. On the other hand, if they are too long, it may fail to capture the "general knowledge" and only stick to particular cases.

Decision Trees

A Decision Tree is a supervised learning predictive model that uses a set of binary rules to calculate a target value. It is used for either classification (categorical target variable) or regression (continuous target variable). Hence, it is also known as CART (Classification Regression Trees). I build a binary decision tree, where at each node we need to make a decision where we split the data using the rule "variable j value"

How do we know what the best split is?

I test all possible combinations of variable and of value, and then choose the split which splits the data in the most egalitarian fashion

How do we know if the data is being well split?

I try to minimize metrics such as entropy or maximize metrics such as the Gini coefficient. The objective is to reduce the original entropy down to zero.

1.2 Research question and methodology

Collecting different Corpora of Debates and Human interactions and, trying to find similar words sequences, NGRAMS that are used in different sentences and grouping those sentences together with Unsupervised Clustering techniques. After Clustering, Finding the group that has words that are closely related to counterargument and insults.

Process

1: Collecting Datasets

Collected Datasets and put them into single repository.

Datasets used

Intelligence Squared Debate Dataset

This dataset contains a collection of transcripts and metadata for debates from the series "Intelligence Squared Debates" (IQ2), held in the US from September 2006 to September 2015. For each debate, the transcript of each turn is given, along with information such as voting results pre- and post-debate, and audience reaction markers. There are 108 debates with an average of 117 utterances per debate.

MPC A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse.
This DataSet is collected from multi-party online conversations in a chat-room environment

Cornell Movie–Dialogs Corpus

This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts:

220,579 conversational exchanges between 10,292 pairs of movie characters

2: Extracting sentences of conversation from corpuses and creating a merged Dataframe from all the documents.

3:Created NGRAMS of each row of extracted sentences to 2Grams and 3 grams, In order to capture sequences that leads to AD hominem, and created a vocabulary of max features upto 100000.

4:TFIDF Vectorizing:TF-IDF is way to find important words in a sentence, by checking how much times a word appears and its relevance in whole document. TFIDF vectorization is building a big Vector consisting of each unique word,Each sentence is a vector, the sentences entered are matrix with 3 vectors. In each vector the numbers (weights) represent features tf-idf score.

5: Trained a classifier to perform a multi-class classification of documents.

Unsupervised Classification:

KMEANS : I clustered all sentences into 50 clusters, with ground truth,seeing the words are not overlapping a lot, and the clusters grouped together are making sense,and did a fixed random seed,so every time for predictions, Clusters do not change their positions.

6:Selected Clusters which contained words making more sense, that could contain words for AD hominem Fallacy.

7:Predicted the systemized movie dataset and filtered with the selected clusterd

8:Manually labeled all the rows of the filtered predicted moviesdataset to check the results.

9:Further trained a Decision tree model to do a Supervised Classification over the manually labeled dataset.

1.3 Experimental results

Unsupervised Classification K means Clusters

```
[ ] for i in range(true_k):
    print(i)
    for ind in order_centroids[i, :10]:
        print(terms[ind])
```

```
be all right
it all
be all
are you
ll be all
8
of course
of course not
course not
of course you
course you
course like
of course it
course it
course did
course noti
9
you said
you said you
said you
thought you said
what you said
thought you
you said it
you said that
what you
said it
10
do you
```

Selected Clusters Cluster 17 and Cluster 38

```
[ ] cluster17
```

```
'fuck are you',
'shut the fuck',
'the fuck up',
'shut the',
'fuck up',
'are you',
'who the fuck',
'the fuck out',
'the fuck you',
'fuck out',
'get the fuck',
'who the',
'fuck you',
'fuck out of',
'get the',
'fuck is going',
'the fuck do',
'fuck do',
'fuck is that',
'is going on',
'out of',
'you talking',
'fuck do you',
'the fuck did',
'fuck did',
'going on',
'fuck was',
'the fuck was',
'is going',
'the fuck outta',
'fuck outta',
'you doing',
'are you talking',
'where the fuck',
'out of here',
'of here',
'are you doing',
'is that',
'talking about',
'how the fuck',
'you talking about',
'fuck is this',
'fuck outta here',
'fuck did you',
'fuck is he',
```

```
cluster38.append(terms[ind])

[ ] cluster38
'much is',
'who the hell',
'and is that',
'that bad',
'what you re',
'supposed to',
'is that for',
'that problem',
'what you want',
'talk is',
'you re',
'that how you',
'that which',
'that like',
'when is',
'who the',
'is that where',
'is that to',
'is that in',
'that why you',
'how much is',
'that all you',
'you want',
'and is',
'is that really',
'what think it',
'that yes',
'gun is',
'is that an',
'that other',
'that what think',
'how many',
'told you',
'that it',
'that one',
'that for',
'point is that',
'that anyway',
'you are',
'and why',
'that what it',
'is that my',
'why you',
'how you',
'what it',
'what you call',
'you call it',
'that funny',
'that what re',
'now who',
'that old',
'that what said',
'all you',
'what book',
'what changed',
```

Labeled Dataset

```
[60] labeledmovies.iloc[480]
```

```
Unnamed: 0          125050
Unnamed: 1              1
paragraphs    Fixing your makeup with a phoneuhWho the fuck are you calling?
pre          17
Name: 480, dtype: object
```

```
labeledmovies.iloc[723]
```

```
Unnamed: 0          183603
Unnamed: 1              1
paragraphs    Now she's tellinme what the fuck to do.
pre          17
Name: 723, dtype: object
```

```
[10] labeledmovies.iloc[626]
```

```
Unnamed: 0          157315
Unnamed: 1              1
paragraphs    What the hell is that anyway"something like th...
pre          38
Name: 626, dtype: object
```

Precision

The precision and recall give a better sense of how an algorithm is actually doing, especially when we have a highly skewed dataset or weak labels. If we predict 0 all the time and get 99.5% accuracy, the recall and precision both will be 0. Because there are no true positives. So, We know that classifier is not a good classifier. When the precision and recall both are high, that is an indication that the algorithm is doing very well.

Unsupervised Classification Precision Score

```
[48] import sklearn
      print(sklearn.metrics.precision_score(labeledmovies['Unnamed: 1'], labeledmovies['preds']))

0.6082383873794917
```

Supervised Classification over Labelled Data with Decision Trees

```
print(sklearn.metrics.precision_score(y_test, preds))

0.8990228013029316
```

1.4 Concluding remarks

The aim for the the project was to Classify text for AD hominem, Though the problem I faced was to exactly understanding and finding corpuses that can have AD hominem fallacy, In this report I mostly focused on Abusive counter arguments or interrogation that are also classified as AD hominem The project can be further improved by creating a very Huge corpus and data set, in order to group together more sequences of words that can have essence of counterargument but in context of insult and not in context of the main topic, and further use pre trained models like BERT or training custom model on Huge corpus with neural network for context word emebedding and then perform various unsupervised classification techniques over those embedding.

1.4.1 References and Citations

Knowledge Extraction and Information Retrieval, alfo ferrara
Conversational flow in Oxford-style debates. Zhang, J., Kumar, R., Ravi, S., Danescu-Niculescu-Mizil, C. (2016). Conversational flow in Oxford-style debates. MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse. Shaikh, S., Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Taylor, S. M., Webb, N. (2010, May). MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse. In LREC
