Weird News Classification & Ranking

IRE-Project-Report

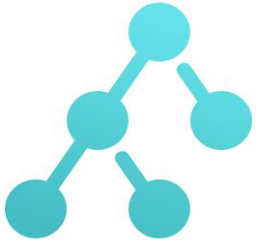
Group: 13
Mentor: Vijayasaradhi Indurthi

What is a weird news?

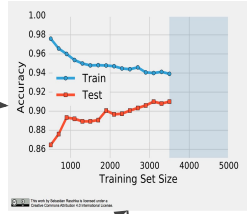
These are the kind of news wich after reading them induces a sense of disbelief or alienation.

FLOW OF THE PROJECT

annotation

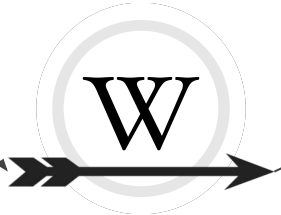


classification



Training

extraction



Word to vec

Ranking

Classification

Figuring out whether a news is weird or not.

Classification

- Classification of the news i.e. whether it is weird or not.
- Comparison with existing Machine Learning techniques to provide a survey of the performance of various models for this task.
- Providing a novel Deep Learning architecture for this classification task.

Classification

Classifiers used:

- ➔ Naive bayes
- ➔ support vectors machines
- ➔ Random forest classifiers
- ➔ Gradient boosting classifiers
- ➔ Ada boosting classifiers
- ➔ Convolutional neural networks
- ➔ Decision tree classifier
- ➔ 3-layered-Perceptron
- ➔ LSTM
- ➔ Auto ML

Experimental analysis of various models for the classification task

Method	Precision	Recall
Naive Bayes	80	78
SVM		
Random Forest Classifier	77	64
Gradient Boosting Classifier	79	79
Ada Boosting Classifier	79	79
Decision Tree Classifier	78	79
3-layered-Perceptron	89	89
LSTM	91	91
CNN	90	90
auto-ML		

Data Annotation

We have developed a good quality dataset that can be used as a gold standard for evaluation.

Data Annotation

Data Annotation App

- A simple application in flask for easy and accurate annotation of weird news.
- Its features includes:

  - multiple annotators allowed at once.

  - Count of total annotation remaining.

  - Its allows user to skip or continue the annotation process later on.

  - Color distinction between annotated(Green), marked(Yellow) and unannotated(Red) news.

Data Annotation

Data Annotation Process

- Each person annotated 500 news sample.
- We rated weird news on a scale of 0-3 where 0 being the least weird and 3 as most weird news.
- All the annotations were done by a single annotator in a file to ensure the annotations have consistency.
- All the annotators finished the annotations in <24 hours of starting so that the annotations have maximum consistency for a single annotator.

RANKING

It involves ranking of weird news with the help of classifiers constructed earlier.

RANKING

The Ranking Task:

- Ranking of the weird news as per its weirdness score.
- Providing a way to rank a set of news based on their weirdness scores.
- Provide a metric that can be used to quantify the weirdness of a news.

RANKING

The Weirdness Probability :

- Weirdness Probability is calculated with respect to the dataset which we have developed with annotation.
- We are using the models like Neural Net , Gaussian ‘s Naive Bayes , Support Vector Machine, Random Forest , Ada Boost and Clustering for finding the probability.
- We are applying the Weirdness ranking based on this probability .

Procedure to Rank

Since we do not have very large annotation set for ranking (just 500 samples), we perform a attempt to learn the level of weirdness using the binary-classified samples.

1. We learn a classifier using the binary classification task. This allows the model to predict the probability of a news article being weird or not.
2. We then use that model and to predict the bucket in which the news article belongs.
3. We then test our model based on the gold standard labels for the test samples. We use the RMSE metric to compare the performance of various methods.
4. Gold Standard Labels: For our task we take gold standard label as the average of the annotations provided by the annotators, scaled in 0 to 1 range.

RANKING

The Ranking is briefly classified into 4 parts:

- Weirdness Probability > 75% will have the Weirdness score as 3
- Weirdness Probability > 50% & <= 75% will have the Weirdness score as 2
- Weirdness Probability > 25% & <=50% will have the Weirdness score as 1
- Weirdness Probability > =0% & <=25% will have the Weirdness score as 0

Clustering based Ranking method

We also experiment with a clustering based ranking method. Here is the procedure that we follow for our method:

1. We use the LSTM based model to learn a classifier for the task.
2. We use the output of the intermediate layer to project a new test sample in the vector space of 300 dimensions.
3. We then cluster all the train samples in that vector space using Affinity Propagation based clustering algorithm and label a cluster as {0, 1, 2, 3} based on the majority of the points in the cluster.
4. Now, for any given point in test dataset, we find its cluster and assign that label to the point.

Experimental analysis of various models for ranking task

Method	RMSE
LogisticRegression	0.41
Naive Bayes	0.54
SVM	N/A
Random Forest Classifier	0.24
Gradient Boosting Classifier	0.33
Ada Boosting Classifier	**0.216**
Decision Tree Classifier	0.54
3-layered-Perceptron	0.45
LSTM	0.519
CNN	0.52

Thanks!

Pinkesh Badjatiya (201402002)
Anupam Pandey (20162118)
Nakul Vaidya (201501108)