IRE

Major Project - Weird News

2nd Deliverable

Introduction

Figuring out the weirdness score and ranking weird or odd news stories

In the Second phase of the project, we will build a classifier to classify a given news article into Bizarre news or conventional news.

In this we will experiment with some of the following features:

We have used 25% for Testing set and 75% for Training set.

Feature Engineering

- Handcrafted features and Linguistic Features::
  - Title length
  - Number of nouns
  - Number of stop words
  - Number of verbs
  - Frequency of co-occurring words
  - word count
  - Average word length
  - ELIPSIS PRESENT OR NOT
  - ! present

- ? present
- : present
- ‘ present
- Number of common verbs in each of 2 classes
- Number of common nouns in each of 2 classes
- Frequent Top words presence OF NORMAL News
- Frequent Top words presence OF WEIRD News
- DICT- Source of the information (url)
- POSSESSIVE WORD Count

● Linguistic Features:

- POS tags (already used nouns, verbs count)
- TF-IDF

The following classification methods will be used along with the above mentioned features:

● Support Vector Machine (SVM)

● Decision Tree

● SVM

● Random Forest

● Logistic Regression

● Deep Neural Network

Using TF-IDF Vector

Model	**Training Score**	**Testing Score**
Random Forest	0.912355174338	0.792919347075
Neural Network	0.8030	0.79808323489284683

using

NOUN,VERB,STOPWORDS,WORDCOUNT,AVGWORDLENGHT,ELIPSIS,EXCLAMATION,QUESTION,COLON,QUOTES,NCV,WCV,NCN,WNN,LABEL

Without TF-IDF Vector, with hand-crafted feature (first 14 features)

Model	Training Score	Testing Score
Random Forest	0.867079005718	0.776957008822
Neural Network	0.7965	0.79316622321803665
Decision Tree	0.87140856576	0.757666993418
Logistic Regression	0.786229431672	0.782033328665
SVM	0.784513945618	0.781718246744

using

NOUN,VERB,STOPWORDS,WORDCOUNT,AVGWORDLENGHT,ELIPSIS,EXCLAMATION,QUESTION,COLON,QUOTES,NCV,WCV,NCN,WNN,NW,WW,LABEL

Model	Training Score	Testing Score
Random Forest	0.909265958688	0.79645707884
Neural Network	0.8175	0.8113009382272498
Decision Tree	0.915275994865	0.773596134995
Logistic Regression	0.808950869413	0.803073799188
SVM	0.807305403198	0.8017784624

using

NOUN,VERB,STOPWORDS,WORDCOUNT,AVGWORDLENGHT,ELIPSIS,EXCLAMATION,QUESTION,COLON,QUOTES,NCV,WCV,NCN,WNN,NW,WW,DICT,LABEL

Model	Training Score	Testing Score
Random Forest	0.909265958688	0.79645707884
Neural Network	0.8142	0.8082551463213542
Decision Tree	0.915275994865	0.773596134995
Logistic Regression	0.808950869413	0.803073799188
SVM	0.807305403198	0.8017784624

using

NOUN,VERB,STOPWORDS,WORDCOUNT,AVGWORDLENGHT,ELIPSIS,EXCLAMATION,QUESTION,COLON,QUOTES,POSSESSIVENESS,NCV,WCV,NCN,WNN,NW,WW,DICT,LABEL

Model	Training Score	Testing Score
Random Forest	0.911798342864	0.798872706904
Neural Network	0.8151	0.80843019183318732
Decision Tree	0.917574979578	0.774646408066
Logistic Regression	0.810117866729	0.804999299818
SVM	0.80956937799	0.803738972133

So , our best model is with

Testing score: 0.8113

Training score: 0.8175

Features
NOUN,VERB,STOPWORDS,WORDCOUNT,AVGWORDLENGHT,ELIPSIS,EXCLAMATION,QUESTION,COLON,QUOTES,NCV,WCV,NCN,WNN,NW,WW,LABEL

Model: Neural Network with 3 layers.