SCIENTIFIC MISINFORMATION DETECTION IN NEWS MEDIA

1. Introduction

In this project we have done the linguistic analysis of news articles in order to detect exaggeration in them while representing the scientific information .

We also try to automate this process by identified features

The dataset which we used for the whole proces: https://figshare.com/articles/InSciOut/903704.

2. Various measures of Exaggeration:

Advice Exaggeration

Implicit Advice : These findings suggest that mid-late childhood may be the best bet for childhood obesity prevention .
Explicit Advice **:** I think we now have enough evidence to say that pulse oximetry screening should be incorporated into everyday clinical practice

Strength of Statement

Low Strength Of Statement : Drinking wine is closely related to cancer .
High Strength Of Statement **:** Drinking wine detoriates body and leads to cancer .

Sample Exaggeration

Studies conducted on particular category and generalized over the whole population is sample exaggeration .

3. Text Preprocessing:

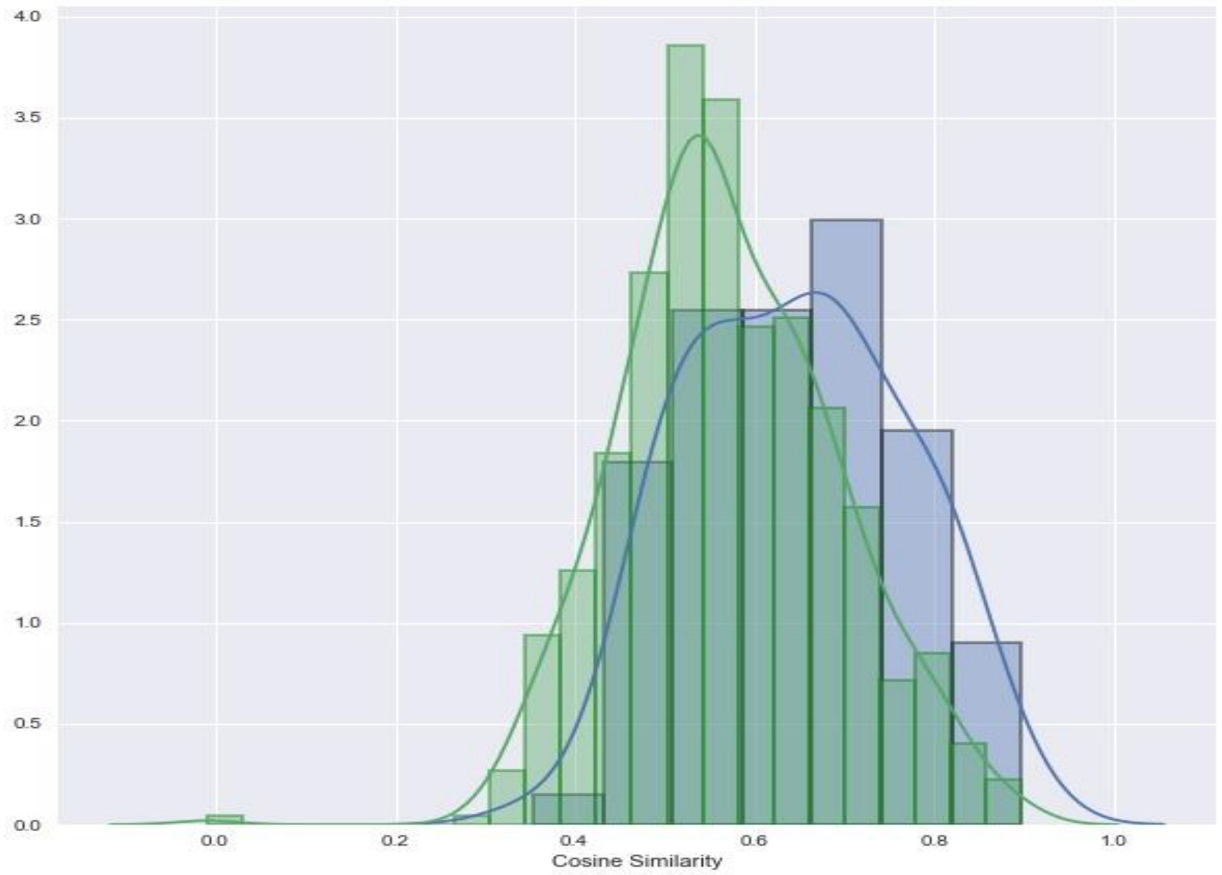
Given dataset contains Press releases and News articles in docs format therefore all docs file were converted into txt format as NLTK tools can be easily applied only txt format files. Then text cleaning was done in which first text is tokenized and all punctuation marks and stop words were removed. Further stemming is done to stem 'snowball' stammers is used. This whole process is done for news articles as well as for press releases.
Further cleaned corpus were created for Press releases and news articles separately.

4. Maxent Model Implementation :

For implementing maxent model first document term matrix is required to be created. We have used 298 high emotion words as vocabulary which is used for creating document term matrix. For each document a vector was created which contain frequency of these high emotion words therefore each document is represented as 298*1 vector.
Now, using these document term matrix maxent model is trained and tested. For training model we have used 70% of news articles as training data and 30% data as test data.

5. Document to vector conversion :

For converting documents into vectors we first created vocabulary containing all press releases and news articles. Then gensim model is used for converting each document into a 300*1 vector.
Then to test relationship between news articles and press releases we computed cosine similarity of news article and corresponding press release. Our hypothesis was that for exaggerated news articles cosine similarity will be less as compared to non-exaggerated articles. Working on that we have observed that hypothesis seems to be true that's why we proceed to consider these doc2vec vectors as features for training our supervised and unsupervised learning models.



6. Supervised Learning Models:

Further for training supervised learning models we have used our 70% news articles as training dataset and rest 30% for test dataset. Now many supervised learning Classifiers were implemented including Logistic regression, SVM, XGBoost, Bagging Classifiers, Random Forest, KNN etc. We implemented our classifiers on two sets of features one set contains only news articles vectors as features and another set which contain concatenated vectors of news article and corresponding press release.

Results obtained are as :

News articles vectors only as features				
Classifier	Advice	Strength of statement	Sample	Total Exaggeration
Logistic Regression	0.77	0.7064	0.87	0.6368
Naive Bayes	0.711	0.7213	0.7468	0.6
Random Forest	0.805	0.741	0.855	0.562
KNN	0.814	0.756	0.86	0.62
XGBoost	0.815	0.751	0.855	0.651
SVM	0.8	0.77	0.87	0.562
Bagging	0.786	0.736	0.855	0.601

Concatenated vectors of news article and press release as features				
Logistic Regression	0.81	0.825	0.94	0.711
Naive Bayes	0.786	0.8	0.88	0.66
Random Forest	0.805	0.756	0.855	0.641
KNN	0.86	0.8	0.855	0.736
XGBoost	**0.845**	**0.84**	**0.92**	**0.791**
SVM	0.805	0.776	0.845	0.582
Bagging	0.835	0.825	0.91	0.726
Logistic Regression	0.81	0.825	0.94	0.711

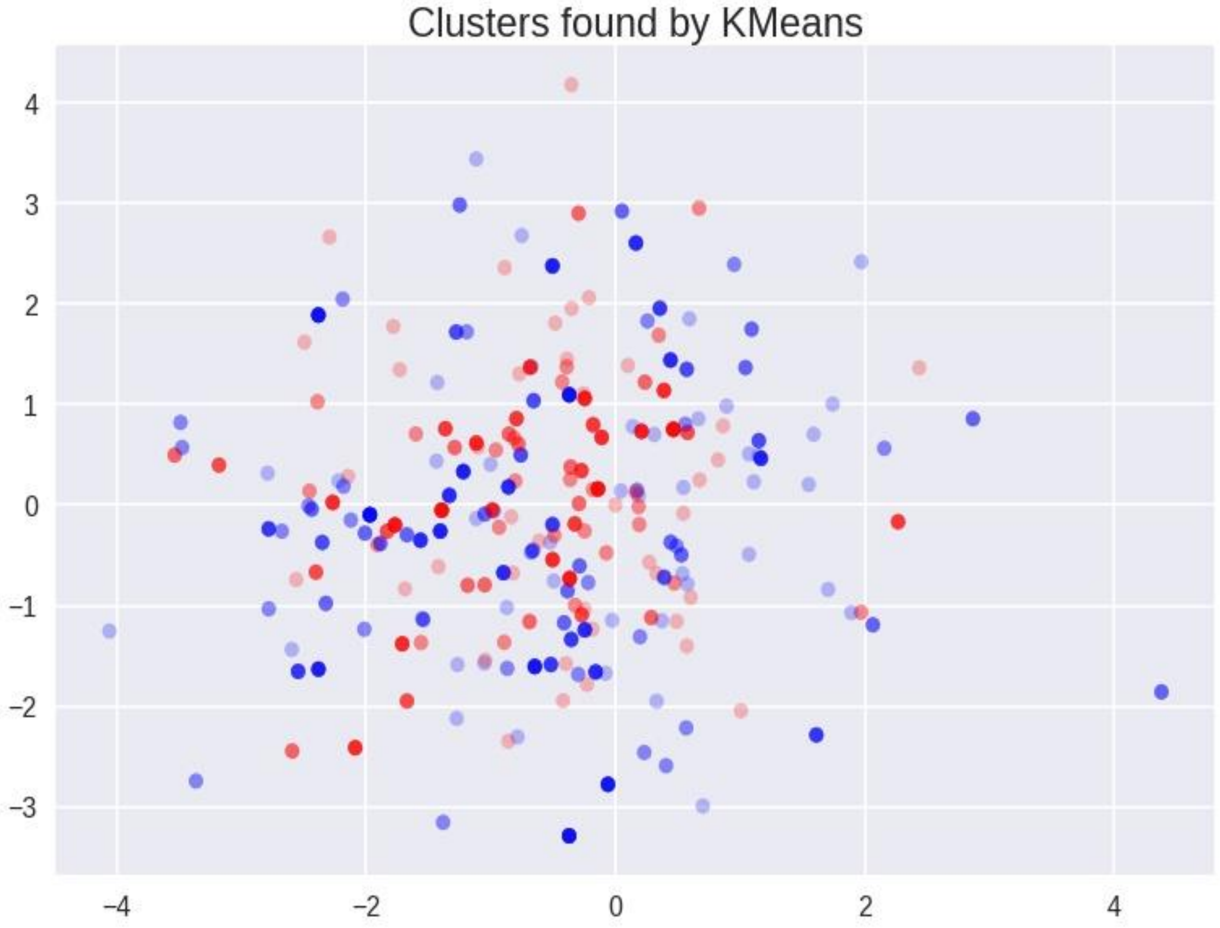
As we can observe from previous results accuracy in case of concatenated vectors is better than only news article vectors as features. Among all models XgBoost, Random forest and Bagging Is best. Therefore considering these doc2vec vectors as features we can classify news articles as Exaggerated or Not Exaggerated.

7. Unsupervised Learning Approach :

We uses clustering algorithms like K-Means, Agglomerative, DBSCAN, HDBSCAN etc.Among them purity of K-Means and Agglomerative observed to be best.

In K-means we computed purity of clusters with K in range 2-15. Using elbow method we obtained that K=10 i best value for obtaining pure clusters.

For clustering we uses doc2vec vectors as features:



References:

1. Link to datset : https://figshare.com/articles/InSciOut/903704.
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054236/
3. http://www.bmj.com/content/349/bmj.g7015
4. https://www.sciencedirect.com/science/article/pii/S0895435616000913
5. https://dl.acm.org/citation.cfm?id=3054270