News Data Analysis

Guide: Dr. D. P. Rana

B.TECH IV

Group 7

Kalp Panwala	U17CO085
Keshav Goyal	U17CO104
Raj Shah	U17CO107
Viren Kathiriya	U17CO113

Motivation

- Fake news is false information presented as news.
- Nowadays, fake news is intentionally written to mislead readers.
- Fake news spreaded over media ecology (from newsprint to radio/television), and recently online news and social media.
- The rapid spread of fake news has the potential for calamitous impacts on individuals and society.

Applications

1. Can stop spread of fake news on social media.
2. Detecting dishonest behavior of retailers.
3. Cannot manipulate elections by detecting Fake News.

Problem Statement

- The prevalence of fake news has attracted increasing attention from researchers to politicians.
- To build a solution that analyse news data i.e. fake news detection using granularity concept.

Objectives

- Detecting phony behaviour of news articles which can make an impact and maintain the social trust.
- Divide the attributes into respective defined granularity ie. Coarse Grained (Topic, Sentence, Document Level features) and Fine Grained (Word Level features).
- Apply Machine Learning techniques to analyse the result.

Literature Review

Authors	Paper Titles	Models Used	Features
Ethan Fast, Bin Binbin Chen, Michael Bernstein(2016)	Empath: Understanding Topic Signals in Large-Scale Text	Empath,LIWC	Text classification, neural network training, 200 in-built features
Qazinian et al. (2011)	Identified tweets in which rumor is endorsed.	Naive Bayes	Content Based, network based, Twitter Specific memes
Gupta et al. (2013)	Analysis of Twitter content during Boston Marathon.	Logistic Regression	Topic engagement, Global engagement, Social reputation, Likability, Credibility
Ning Cao et al. (2020)	A deceptive review detection framework	LDA-BP + TextCNN + SVM	Fine-grained and coarse-grained features

Literature Review

Authors(ref)	Paper Titles	Models Used	Features
Ahmed et al. (2017)	Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques	SVM, LSVM, KNN,DT, Logistic Regression	TF-IDF
Markines et al. (2009)	Social Spam Detection using Adaboost	SVM, Ada-Boost	Tag Spam, TagBlur, DomFp, NumAds, Plagiarism, ValidLinks
S. Chhabra et al. (2011)	Fake and Malicious URL Detection	Naive Bayes, Logistic Regression, DT, SVM RBF, SVM Linear SVM Sigmoid	Grammar, Lexical, Vectors and Static.
Chen, Konroy and Rubin(2015)	Deception Detection and 3 Types of Fakes	NLP, Sentiment Analysis, Big Data	Type A - Serious Fabrication, Type B - Large Scale Hoaxes, Type C - Humorous Fakes

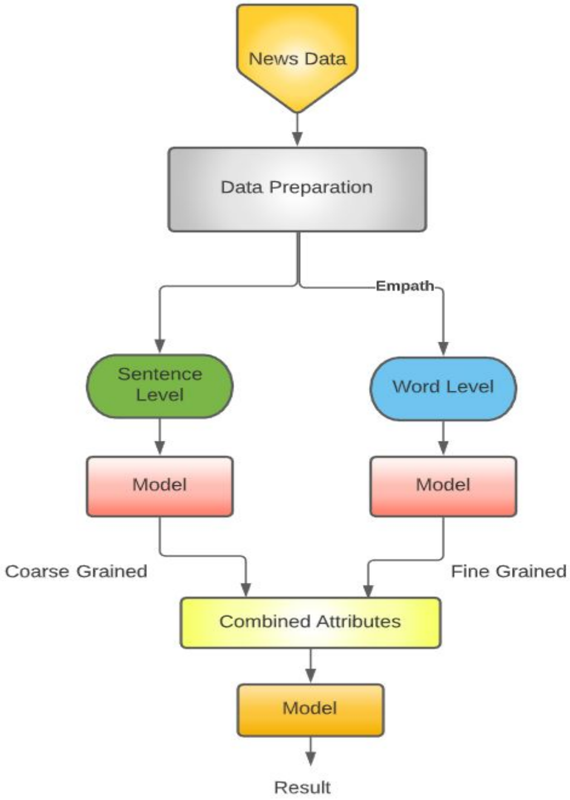
Literature Review

Authors(ref)	Paper Titles	Models Used	Features
P. Lakshmi Prasanna, Dr. Rao (2018)	Text classification using artificial neural networks	ANN, Document conversion, stemming	TF-IDF Matrix from text classification.
Matthew Whitehead, Larry Yaeger(2008)	Sentiment Mining Using Ensemble Classification Models	Bagging, Boosting, single model SVM, K-fold(10) cross validation	Ensemble Classifiers

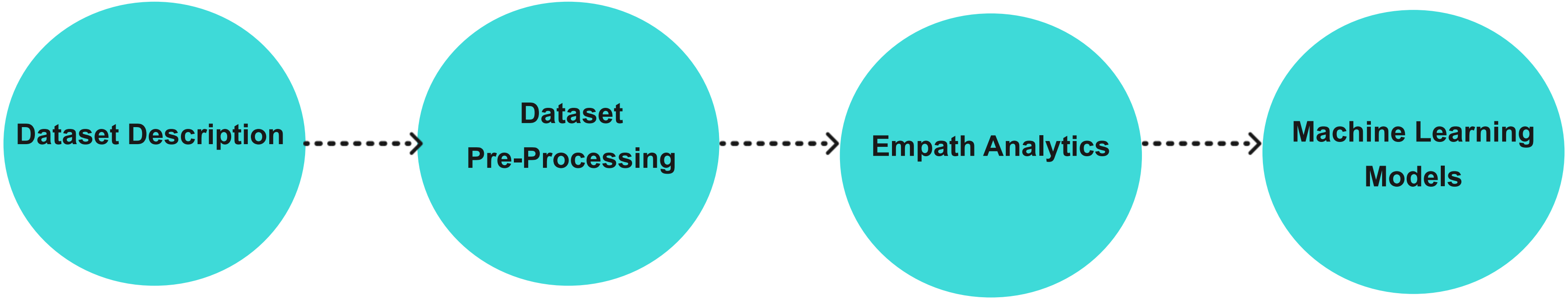
Fine and Coarse Grain Features

- Fine Grained Features
  - The smallest possible meaningful content in a topic model can be a word which defines Fine Grained features.
  - Eg. **Violence** is a attribute with seed words hurt, break, bleed, broken, etc..
- Coarse Grained Features
  - Explicitly defined as overall data in the text which has a tendency to split enough.
  - Eg. War is indeed painful. This sentence indirectly specifies **Violence**.

Proposed Framework



Solution Flow (Fine Grained)



Dataset Description

Dataset Pre-Processing

Empath Analytics

Machine Learning Models

- Dataset consist of 10000 articles.
- The features of the dataset are title, text, subject, date, category.

.

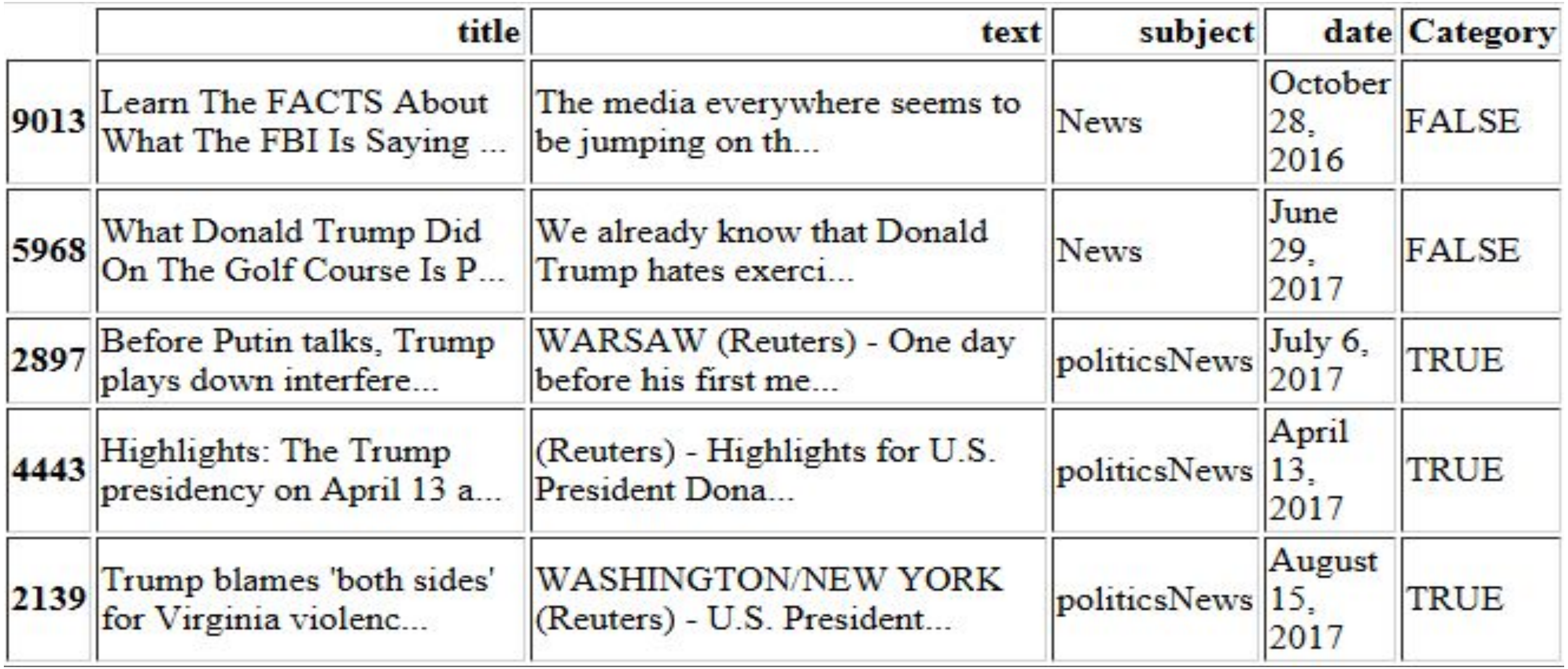
- Lowercasing, Lemmatization, Stop-word removal.
- Missing Value Replacement.
- Text Reduction.
- Text Normalization.

- Tool for analyzing text across lexical categories.
- Classifies into around 200 attributes.

- Train models on various dataset discussed further.

Dataset Analysis

Dataset Source: Kaggle (Click to download.)

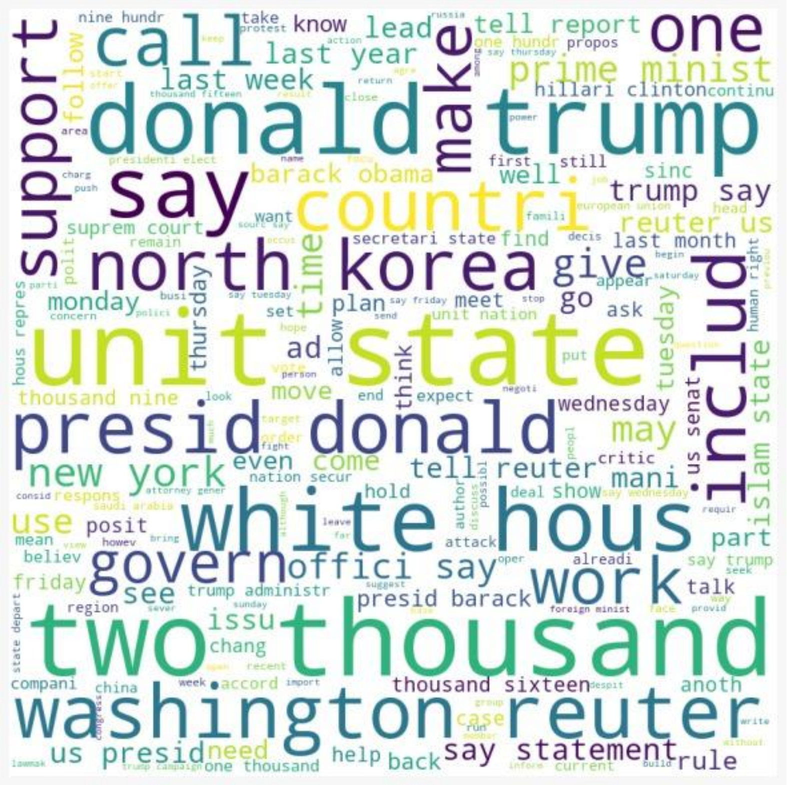


	title	text	subject	date	Category
9013	Learn The FACTS About What The FBI Is Saying ...	The media everywhere seems to be jumping on th...	News	October 28, 2016	FALSE
5968	What Donald Trump Did On The Golf Course Is P...	We already know that Donald Trump hates exerci...	News	June 29, 2017	FALSE
2897	Before Putin talks, Trump plays down interfere...	WARSAW (Reuters) - One day before his first me...	politicsNews	July 6, 2017	TRUE
4443	Highlights: The Trump presidency on April 13 a...	(Reuters) - Highlights for U.S. President Dona...	politicsNews	April 13, 2017	TRUE
2139	Trump blames 'both sides' for Virginia violenc...	WASHINGTON/NEW YORK (Reuters) - U.S. President...	politicsNews	August 15, 2017	TRUE

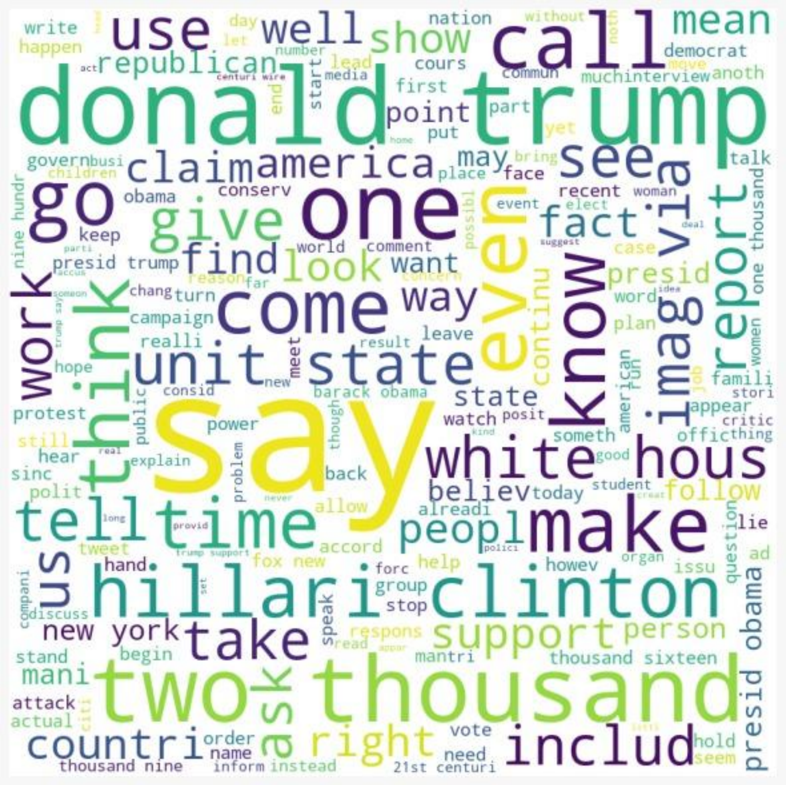
Simulation And Results

- **Word Cloud**
  - **Data Visualization** technique used for representing text data in which the size of each word indicates its frequency or importance.
  - Larger the height of the word, more it ensembles that class.

Analysis Of Results



True News Word Cloud



False News Word Cloud

Simulation and Results

- Fine Grained Features are extracted using Empath.
  - Around 200 features.

- ML Classifiers
  - Logistic Regression.
  - KNN with n=3.
  - SVM.
  - Random Forest.
  - Gradient Boosting Algorithm.

Simulation And Results

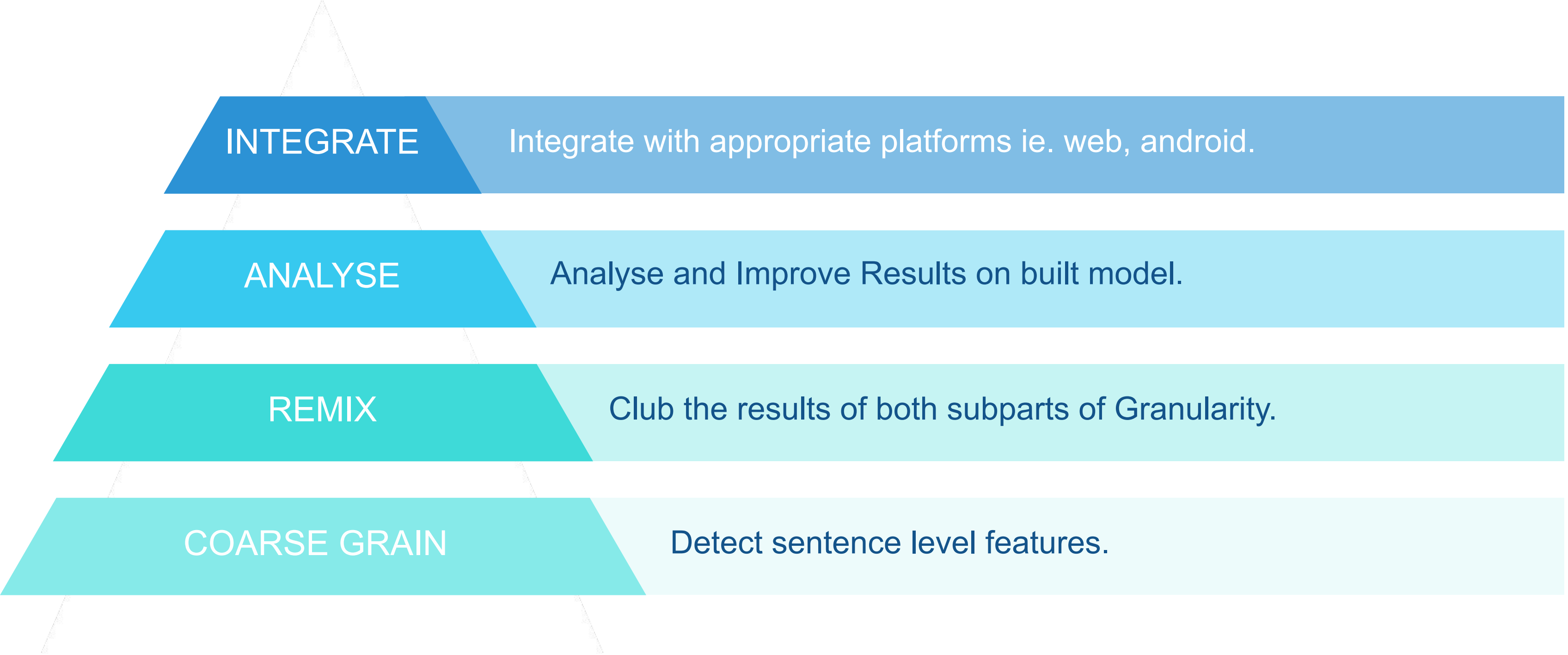
- Evaluation metrics
    - **Accuracy**
    - **F1-Score**
- Result
    - Accuracy and F1-score 92 %.

Conclusion

- Data Preprocessing was a core part along with feature extraction.
- We conclude a part of granularity, ie. Fine Grained on textual news. which drives deeper into word level features for analysis of the required news.

Link to Report :- Click here

Future Works



INTEGRATE

Integrate with appropriate platforms ie. web, android.

ANALYSE

Analyse and Improve Results on built model.

REMIX

Club the results of both subparts of Granularity.

COARSE GRAIN

Detect sentence level features.

References

1. B. Markines, C. Cattuto, F. Menczer, in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (ACM, 2009), pp. 4148.

2. Ning Cao, Shujuan Ji, Dickson K.W. Chiu, Mingxiang He, Xiaohong Sun, supported in part by the Natural Science Foundation of China (No. 71772107, 61502281)

3. V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, 2011), pp. 15891599

4. Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. $1.00 per rt boston marathon prayforboston: Analyzing fake content on twitter.

5. Ahmed, Hadeer Saad, Sherif. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. 127-138. 10.1007/978-3-319-69155-8_9.

6. ASIST 2015, November 6-10, 2015, St. Louis, MO, USA. Copyright l’ 2015 Yimin Chen, Victoria L. Rubin and Niall J. Conroy.

7. S. Chhabra, A. Aggarwal, F. Benevenuto, P. Kumaraguru,in Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (ACM, 2011), pp. 92-101.

8. Empath:Understanding Topic Signals in Large-Scale Text ACM Classification Keywords H.5.2. Information Interfaces and Presentation: Group and Organization Interfaces.

9. Text classification using artificial neural networks: International Journal of Engineering Technology Website: www.sciencepubco.com/index.php/IJET

10. Chaitanya Naik, Vallari Kothari, Zankhana Rana, Document Classification using Neural Networks Based on Words, In: International Journal of Advanced Research in Computer Science,2015.

References

11. Snyder, B., and Barzilay, R. 2007. Multiple aspect ranking using the good grief algorithm. In Proceedings of NAACL HLT, 300307.

12. Schapire, R. E. 2002. The boosting approach to machine learning: An overview

13. Sentiment Mining Using Ensemble Classification Models Conference Paper, January 2008

14. James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. In Mahwah: Lawrence Erlbaum Associates 71 2001.

15. Confusion Matrix-based Feature Selection :Conference Paper. January 2011 by Sofia et al.

16. Amador Diaz Lopez, J., Oehmichen, A., Molina-Solana, M.: Fake News on 2016 US elections viral tweets (November 2016 March 2017), November 2017.

17. Review of Data Preprocessing Techniques in Data Mining : Article in Journal of Engineering and Applied Sciences. September 2017 DOI: 10.3923/jeasci.2017.4102.4107

18. A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM : Journal of Adv Research in Dynamical Control Systems, Vol. 11, 06-Special Issue, 2019

19. W. McKinney, pandas: a python data analysis library, http: //pandas.sourceforge.net

20. Fake News Detection Method Based on Text-Features: IMMM 2019 : The Ninth International Conference on Advances in Information Mining and Management

21. Research on Data Preprocessing in Exam Analysis System : College of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, china

Thank you!!