News Data Analysis

Guide: Dr. D. P. Rana

B.TECH IV

Group 7

Kalp Panwala	U17C0085
Keshav Goyal	U17C0104
Raj Shah	U17C0107
Viren Kathiriya	U17C0113

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
Ning Cao et al. (2020)	A deceptive review detection framework	LDA-BP + TextCNN + SVM	Fine-grained and coarse-grained 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
Jae-Seung Shim et al (2019)	Document Summarization Technique on the Fake News Detection Model	PCA, SVM, Regression, Decision Tree	Lexrankr to get 3 line summary.
Jing Li et. al (2020)	A Survey on Deep Learning for Named Entity Recognition	CNN, LSTM, encoder, Tag Decoder.	Traditional NER, Deep Learning NER with neural nets.

Literature Review

Authors	Paper Titles	Models Used	Features
Ritter et.al (2011)	Named Entity Recognition in Tweets:An Experimental Study	Named Entity Recognition.	Postagging, Shallow Parsers,LDA
Savelieva et.al (2020)	Abstractive Summarization of Spoken and Written Instructions with BERT	Text summarization	NLP,BERT,Neural Network.
Castelo et al. (2019).	A Topic-Agnostic Approach For Identifying Fake News Pages.	SVM, KNN, Random Forest	Morphological Features, Psychological Features, Readability Features, Web-Markup Features.
Kuai Xu et al. (2020)	Detecting Fake News Over Online Social Media via Domain Reputations and Content Understanding	LDA Topic Modelling	TF-IDF

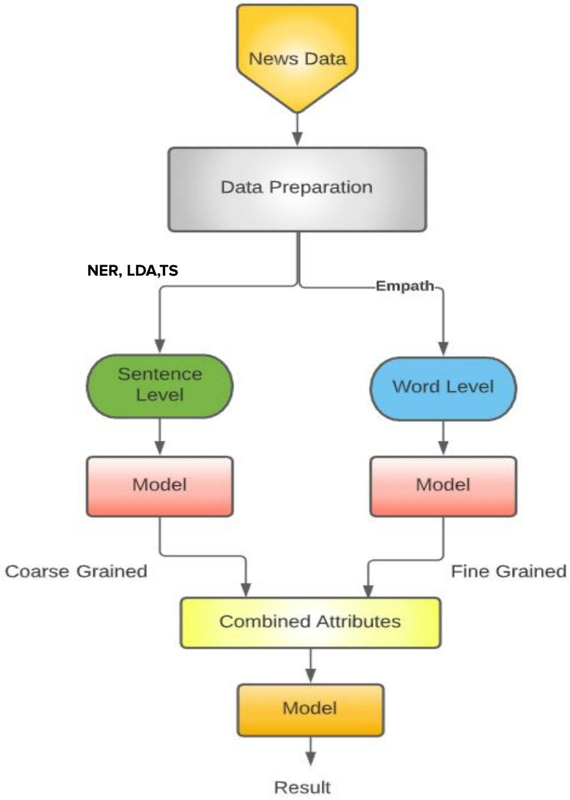
Literature Review

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

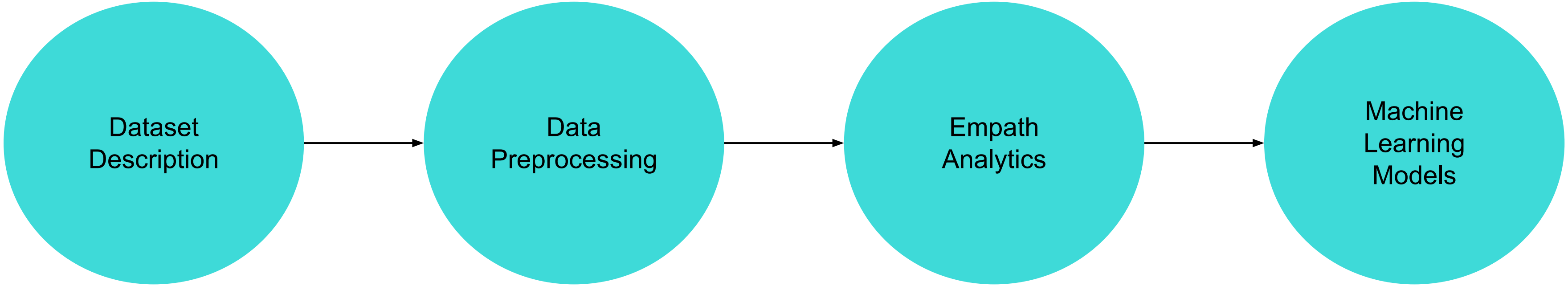
Fine and Coarse Grain Features

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

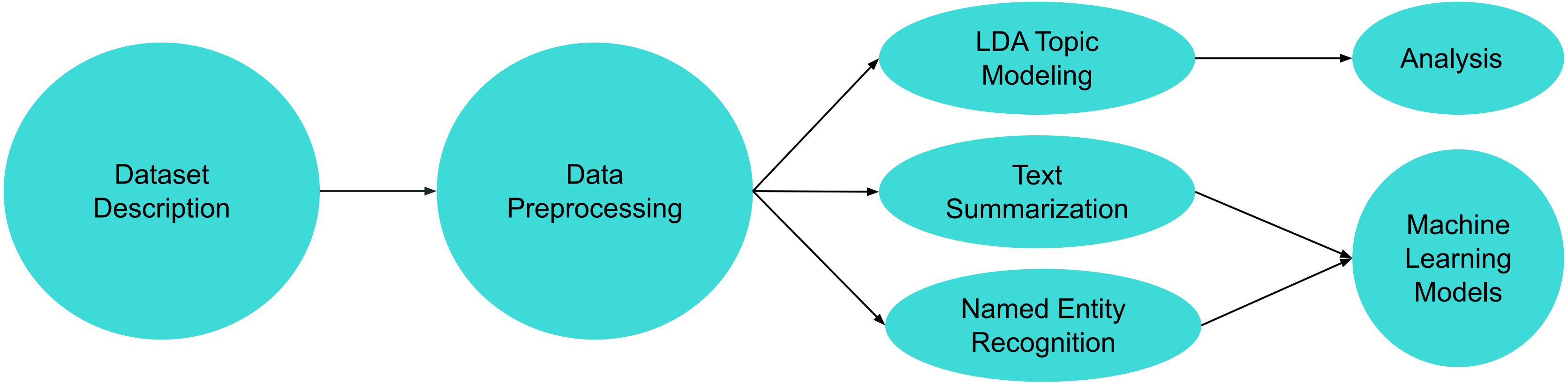
.

- 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.

Solution Flow (Coarse Grained)



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

.

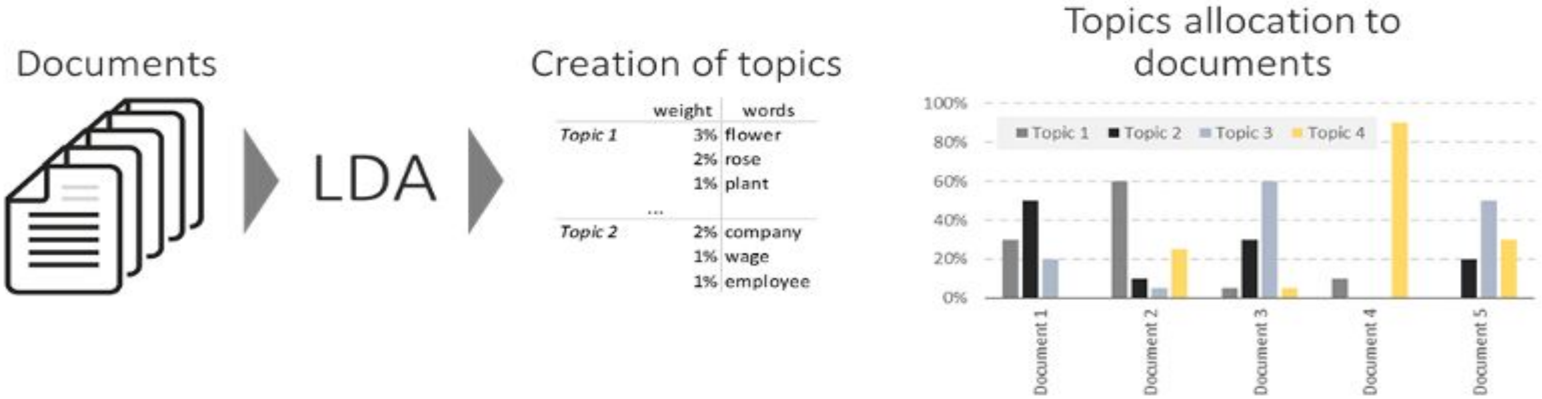
- Lowercasing, Lemmatization, Stop-word removal.
- Missing Value Replacement.
- Text Reduction.
- Text Normalization.

- Classifies sentences into topics.
- Each topic consists of predefined combination of words.

- Train models on various dataset discussed further.

Coarse Grain - LDA

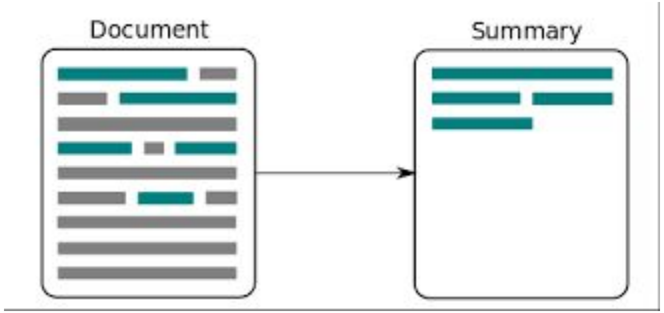
- Latent Dirichlet Allocation - Topic Modeling Technique.
- Statistical modeling technique.
- Builds topic per document model and words per topic model.
- Dictionary and the Corpus are the two main inputs.
- Python’s Gensim package is used for implementation.



	weight	words
Topic 1	3%	flower
	2%	rose
	1%	plant
...		
Topic 2	2%	company
	1%	wage
	1%	employee

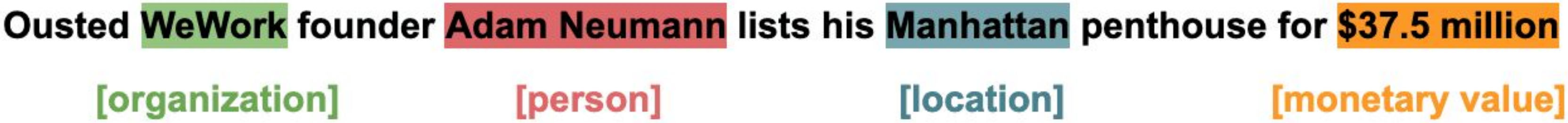
Coarse Grain - Text Summarization

- Reduces long pieces of text.
- Creates a summary of the text having the main points outlined.
- Text Summarization techniques
  - Extractive methods - selecting phrases and sentences from source documents to make the new summary.
  - Abstractive method - generating new phrases and sentences that have the same meaning as the source document.



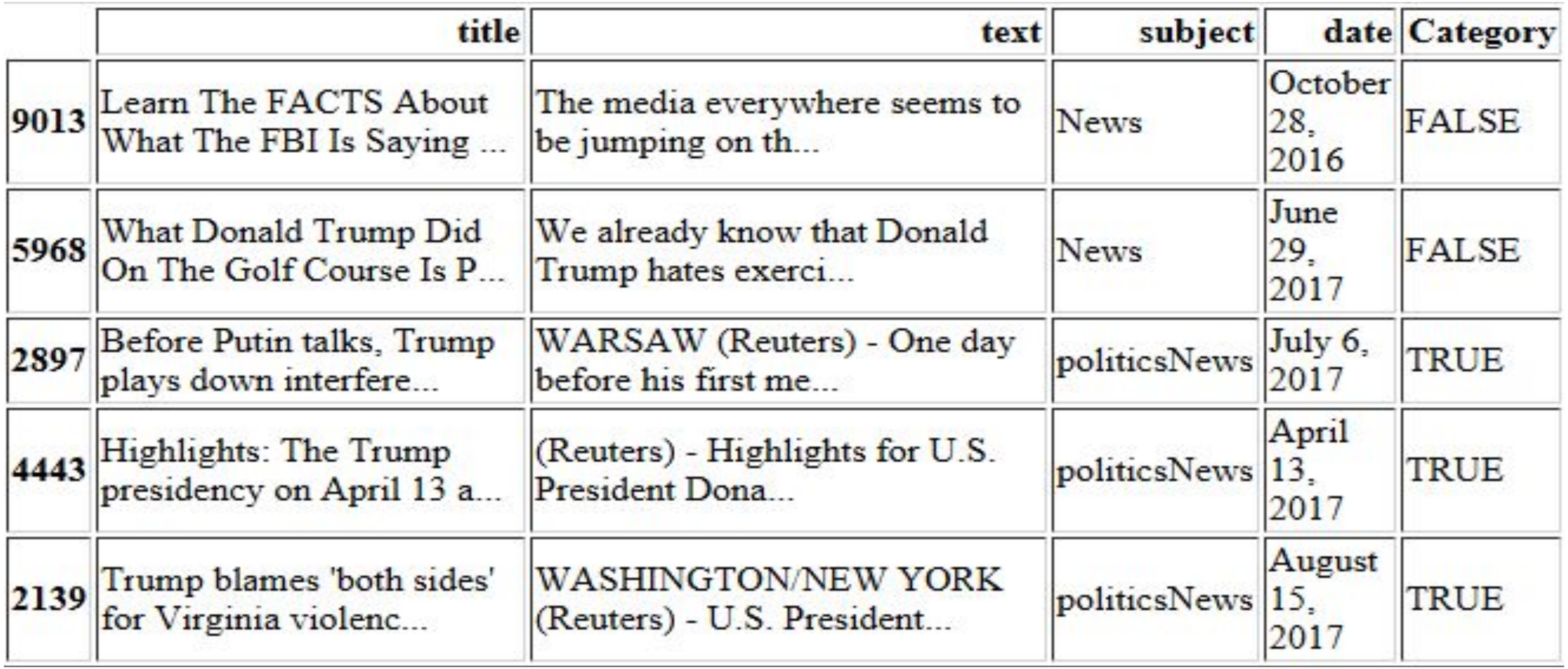
Coarse Grain - NER

- Named Entity Recognition
- Subtask of Information Extraction
  - seek to locate and classify named entities from unstructured text
  - Map them to predefined categories such as person names, organisation, locations etc
- Majorly uses SpaCy for implementation
- Uses large volumes of twitter texts, unstructured data, emails, feeds, etc. to predict named entities in a given corpus or sentence.


Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million
[organization]
[person]
[location]
[monetary value]

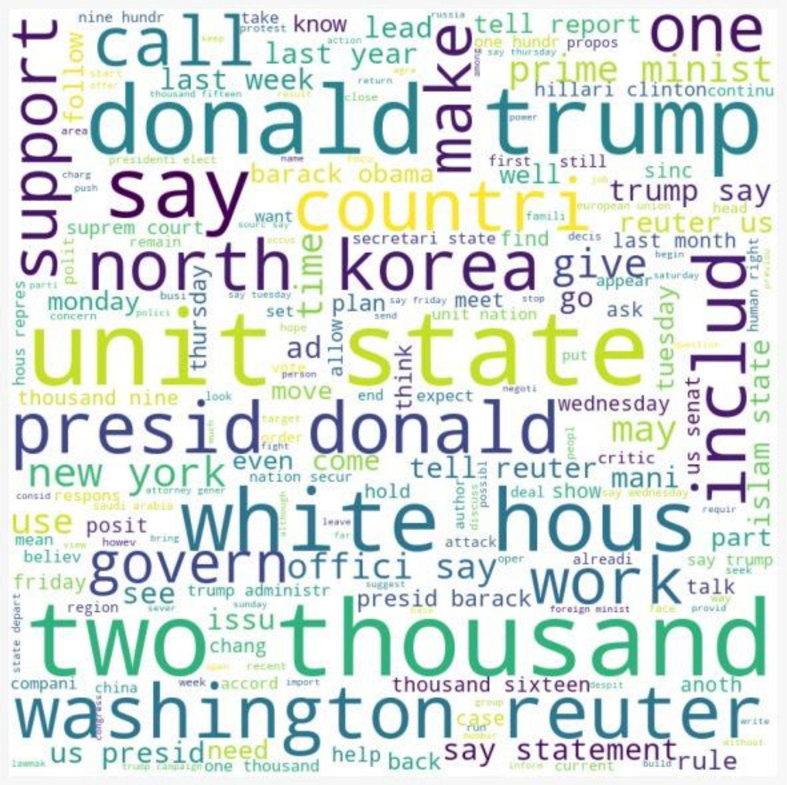
Dataset Analysis

Dataset Source: Kaggle (Click to download.)

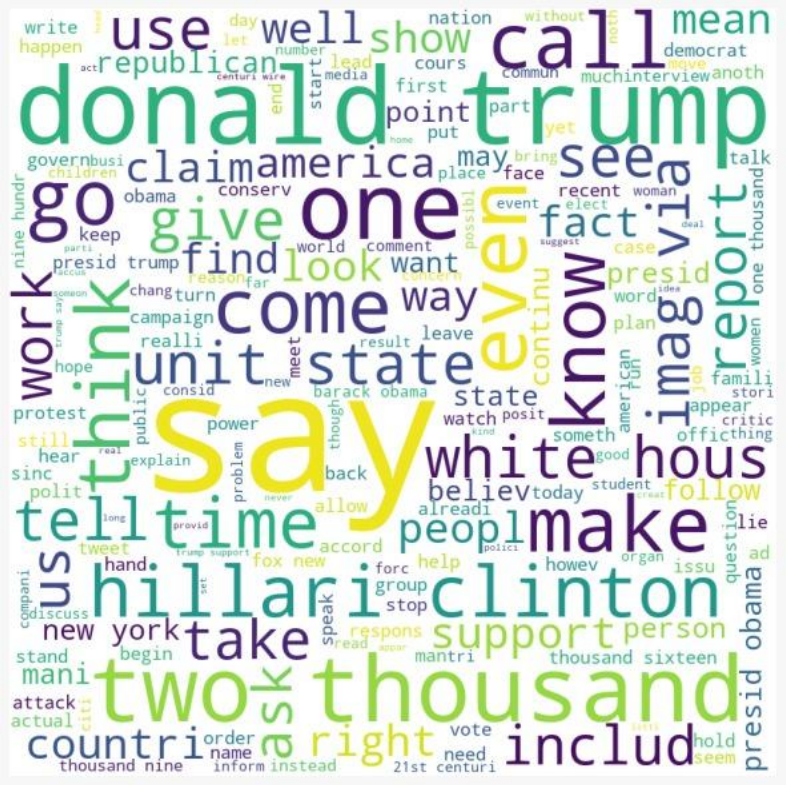


	title	text	subject	date	Category
9013	Learn The FACTS About What The FBI Is Saying ...	The media everywhere seems to be jumping on th...	News	October 28, 2016	FALSE
5968	What Donald Trump Did On The Golf Course Is P...	We already know that Donald Trump hates exerci...	News	June 29, 2017	FALSE
2897	Before 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

Analysis Of Results (Fine Grain)

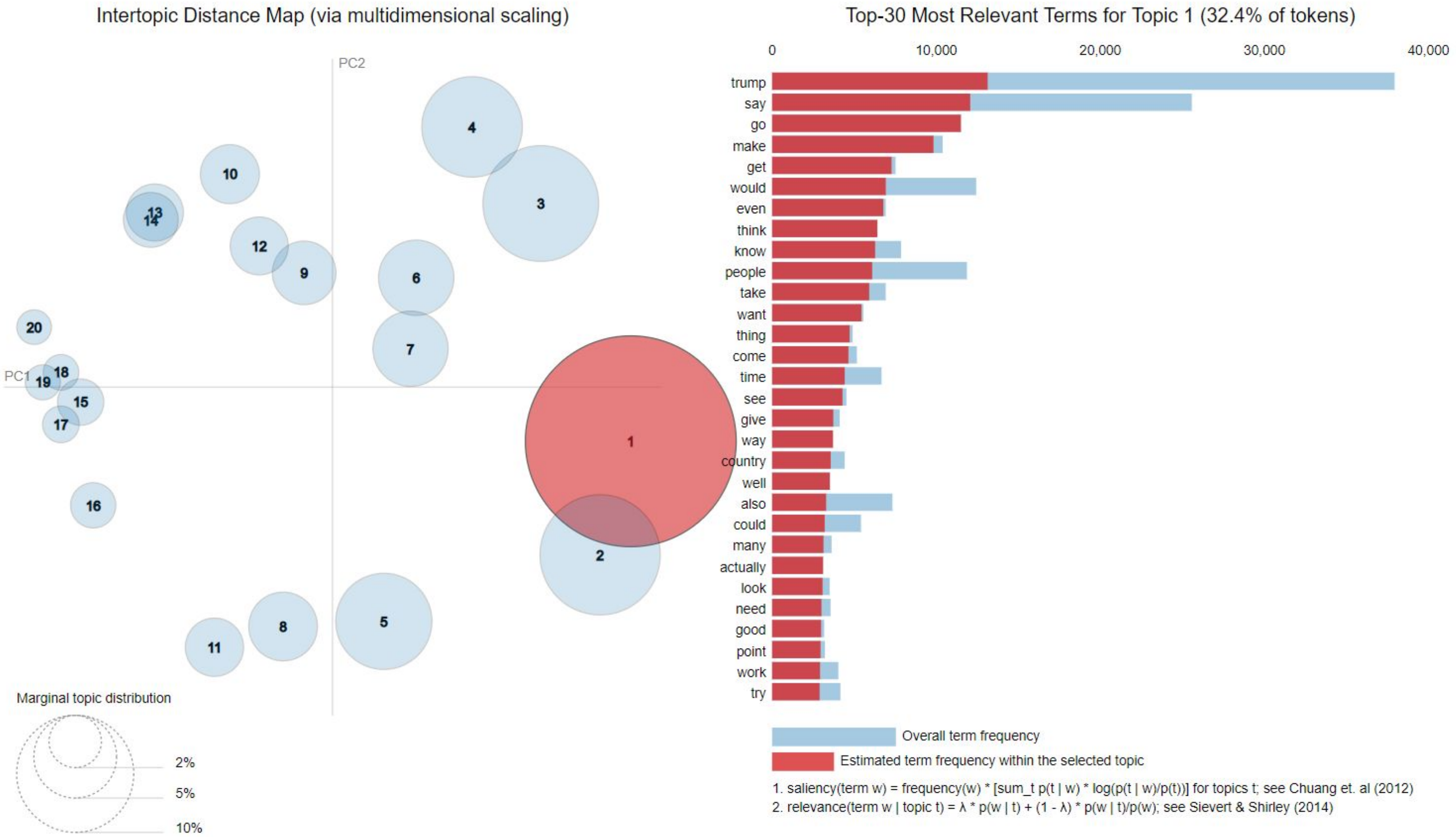


True News Word Cloud



False News Word Cloud

Analysis Of Results (Coarse Grain - LDA)

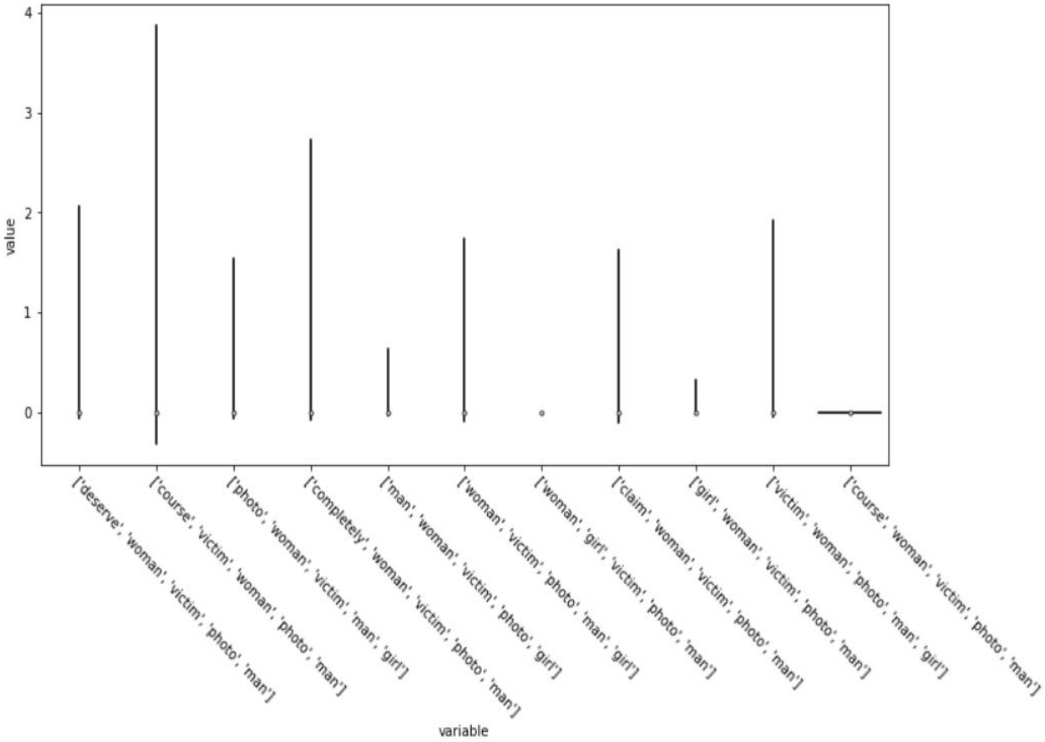


Intertopic Distance Map (via multidimensional scaling)

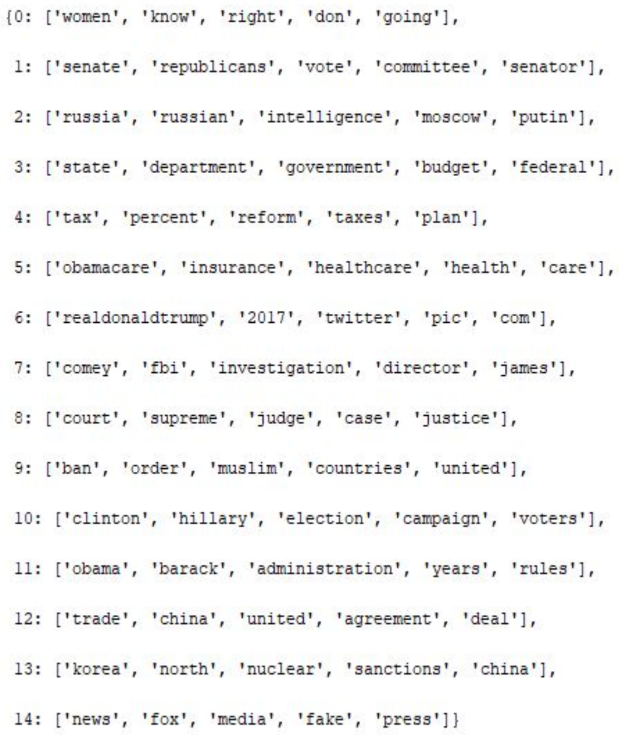
Top-30 Most Relevant Terms for Topic 1 (32.4% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Analysis Of Results (Coarse Grain - TS)



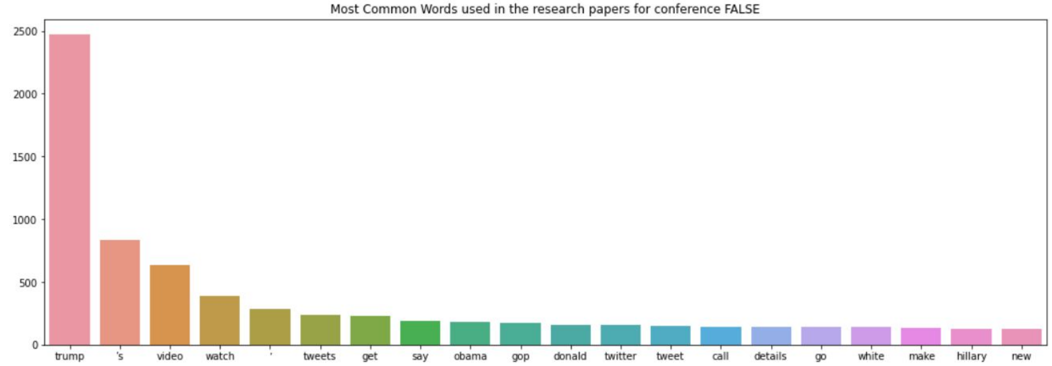
Text Summarization Topics



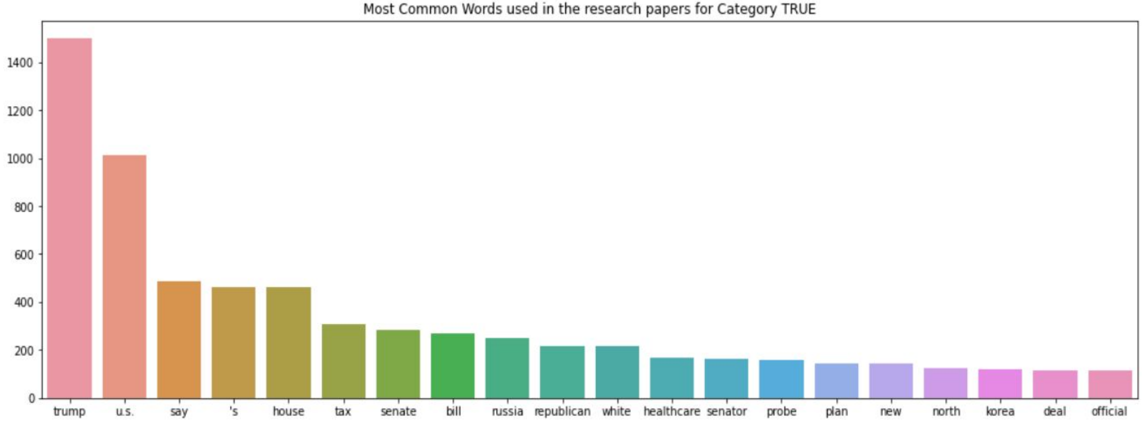
```
{0: ['women', 'know', 'right', 'don', 'going'],

 1: ['senate', 'republicans', 'vote', 'committee', 'senator'],

 2: ['russia', 'russian', 'intelligence', 'moscow', 'putin'],

 3: ['state', 'department', 'government', 'budget', 'federal'],

 4: ['tax', 'percent', 'reform', 'taxes', 'plan'],

 5: ['obamacare', 'insurance', 'healthcare', 'health', 'care'],

 6: ['realdonaldtrump', '2017', 'twitter', 'pic', 'com'],

 7: ['comey', 'fbi', 'investigation', 'director', 'james'],

 8: ['court', 'supreme', 'judge', 'case', 'justice'],

 9: ['ban', 'order', 'muslim', 'countries', 'united'],

 10: ['clinton', 'hillary', 'election', 'campaign', 'voters'],

 11: ['obama', 'barack', 'administration', 'years', 'rules'],

 12: ['trade', 'china', 'united', 'agreement', 'deal'],

 13: ['korea', 'north', 'nuclear', 'sanctions', 'china'],

 14: ['news', 'fox', 'media', 'fake', 'press']}
```

Text Summarization Topics Modeling

Analysis Of Results (Coarse Grain - NER)



NER False Topics



NER True Topics

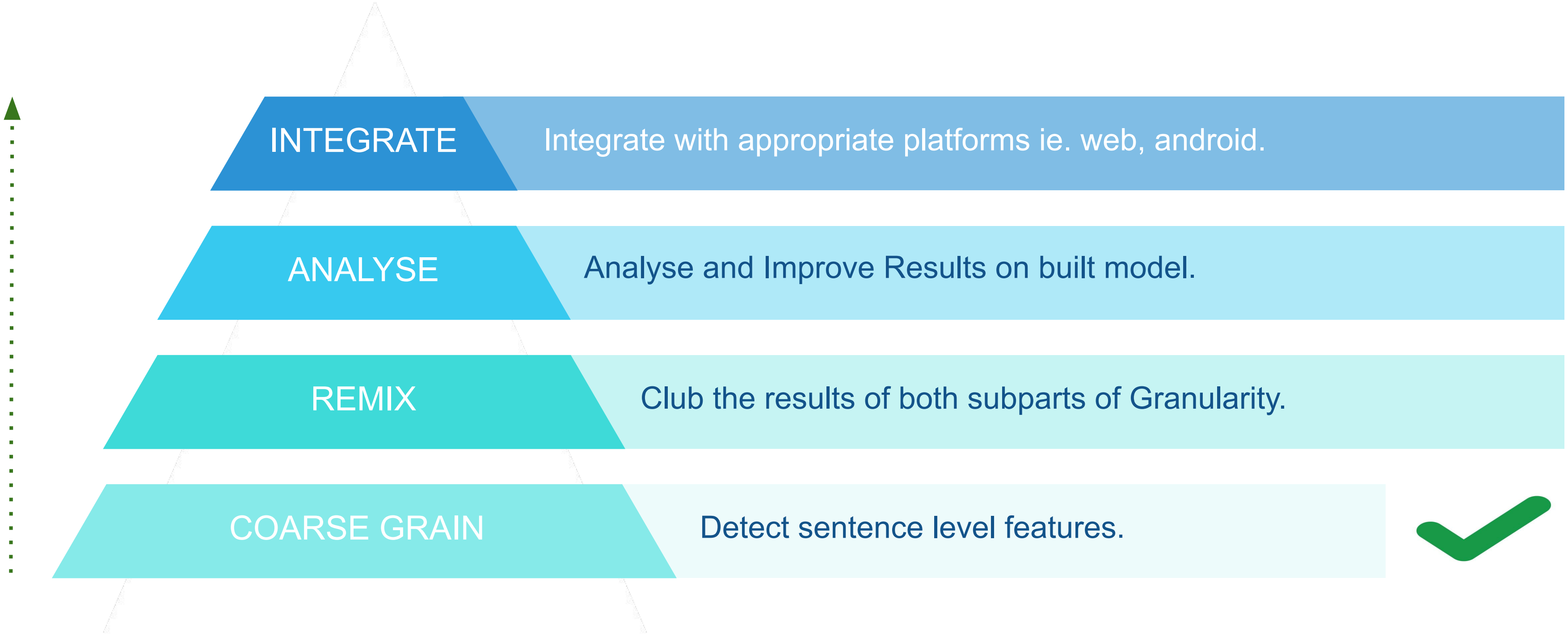
Simulation And Results

- Evaluation metrics
  - **Accuracy**
  - **F1-Score**
  - **Accuracy and F1-Score differed maximum by 0.1%**
- Result
  - Fine Grain: (Empath)
    - Accuracy - 92%
  - Coarse Grain: (NER, Text Summarization, LDA)
    - Accuracy - 95% (max using NER)
- ML Classifiers
  - Logistic Regression.
  - KNN with n=3.
  - SVM.
  - Random Forest.
  - Gradient Boosting Algorithm.

Conclusion

- Data Preprocessing was a core part along with feature extraction.
- We conclude granularity concepts and its implementations, ie. Fine Grain and Coarse Grain on textual 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, "Social spam detection", in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (ACM, 2009), pp. 41-48.

[2] Ning Cao, Shujuan Ji, Dickson K.W. Chiu, Mingxiang He, Xiaohong Sun, "A deceptive review detection framework: Combination of coarse and fine-grained features, Expert Systems with Applications", Volume 156, 2020, 113465, ISSN 0957-4174.

[3] Qazvinian, Vahed, Emily Rosengren, Dragomir R. Radev and Q. Mei. Rumor has it: Identifying Misinformation in Microblogs., in Proceedings of the Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, 2011), pp. 1589-1599

[4] Gupta, Aditi Lamba, Hemank Kumaraguru, Ponnurangam. (2013). "1.00 per RT Boston Marathon PrayForBoston: Analyzing fake content on Twitter" eCrime Researchers Summit, eCrime. 1-12. 10.1109/eCRS.2013.6805772.

[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] Conroy, Nadia Rubin, Victoria Chen, Yimin. (2015). "Automatic Deception Detection:Methods for Finding Fake News." Conference: ASIST 2015: St. Louis, MO, USA.

[7] Chhabra S, Aggarwal A, Benvenuto F, Kumaraguru P (2011) "Phi.sh/social: the phishing landscape through short urls", In: Annual collaboration, electronic messaging, anti-abuse and spam conference (CEAS), Perth, 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] Prasanna, P. Rao, Dr. (2018), "Text classification using artificial neural networks", International Journal of Engineering and Technology(UAE). 7. 603-606. 10.14419/ijet.v7i1.1.10785.

[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.

[11] Snyder, B., and Barzilay, R. 2007. "Multiple aspect ranking using the good grief algorithm" In Proceedings of NAACL HLT, pp. 300-307.

[12] Schapire R.E. (2003) "The Boosting Approach to Machine Learning: An Overview" In:Denison D.D., Hansen M.H., Holmes C.C., Mallick B., Yu B. (eds) Nonlinear Estimation and Classification. Lecture Notes in Statistics, vol 171. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21579-29

[13] Whitehead, Matthew Yaeger, Larry. (2008). "Sentiment Mining Using Ensemble Classification Models", Innovations and Advances in Computer Sciences and Engineering. 509-514. 10.1007/978-90-481-3658-289.

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

References

[15] Visa, Sofia Ramsay, Brian Ralescu, Anca Knaap, Esther. (2011), "Confusion Matrix-based Feature Selection" CEUR Workshop Proceedings. 710.120-127.

[16] Oehmichen, Axel Hua, Kevin Lopez, Julio Molina-Solana, Miguel Gómez-Romero, Juan Guo, Yike. (2019). "Not All Lies Are Equal. A Study Into the Engineering of Political Misinformation in the 2016 US Presidential Election." IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2938389.

[17] Bhaya, Wesam, "Review of Data Preprocessing Techniques in Data Mining" Journal of Engineering and Applied Sciences. 12. 4102-4107. 2017.

[18] Tijare, Poonam, "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] Mckinney, Wes. (2011). "pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer", http://pandas.sourceforge.net.

[20] Drif, Ahlem Ferhat Hamida, Zineb Giordano, Silvia. "Fake News Detection Method Based on Text-Features", proceedings of The Ninth International Conference on Advances in Information Mining and Management, Aug-2019.

[21] Zhu M. (2011) "Research on Data Preprocessing in Exam Analysis System" In: Ma M.(eds) Communication Systems and Information Technology. Lecture Notes in Electrical Engineering, vol 100. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21762-343

[22] K. Xu, F. Wang, H. Wang and B. Yang, "Detecting fake news over online social media via domain reputations and content understanding," in Tsinghua Science and Technology, vol.25, no. 1, pp. 20-27, Feb. 2020, doi: 10.26599/TST.2018.9010139.

[23] Castelo, Sonia Almeida, Thais Elghafari, Anas Santos, Aécio Nakamura, Eduardo Freire, Juliana. (2019). A Topic-Agnostic Approach For Identifying Fake News Pages.

[24] Allahyari, Mehdi Pouriyeh, Seyedamin Assefi, Mehdi Safaei, Saeid Trippe, Elizabeth Gutierrez, Juan Kochut, Krys. (2017). Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Applications. 8. 397-405. 10.14569/IJACSA.2017.081052.

[25] Jae-Seung Shim, Ha-Ram Won, Hyunchul Ahn. (2019). A Study on the Effect of the Document Summarization Technique on the Fake News Detection Model. Journal of Intelligence and Information Systems, Vol -25 No-3, 2019.

[26] li, Jing Sun, Aixin Han, Ray Li, Chenliang. (2020). A Survey on Deep Learning for Named Entity Recognition, IEEE Transactions on Knowledge and Data Engineering. PP.1-1. 10.1109/TKDE.2020.2981314.

[27] Alan Ritter, Sam Clark, Mausam, Oren Etzioni, Named Entity Recognition in Tweets: An Experimental Study, in the Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing Pp 1524-1534, July-2011.

[28] Savelieva, Alexandra Au-Yeung, Bryan Ramani, Vasanth. (2020). "Abstractive Summarization of Spoken and Written Instructions with BERT."

[29] Bíró I., Szabó J. (2009) "Latent Dirichlet Allocation for Automatic Document Categorization." In: Buntine W., Grobelnik M., Mladeni D., Shawe-Taylor J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science, vol 5782. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04174-728.

Thank you!!