News Data Analysis

B.TECH IV

Group 7

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

Guide: Dr. D. P. Rana

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.

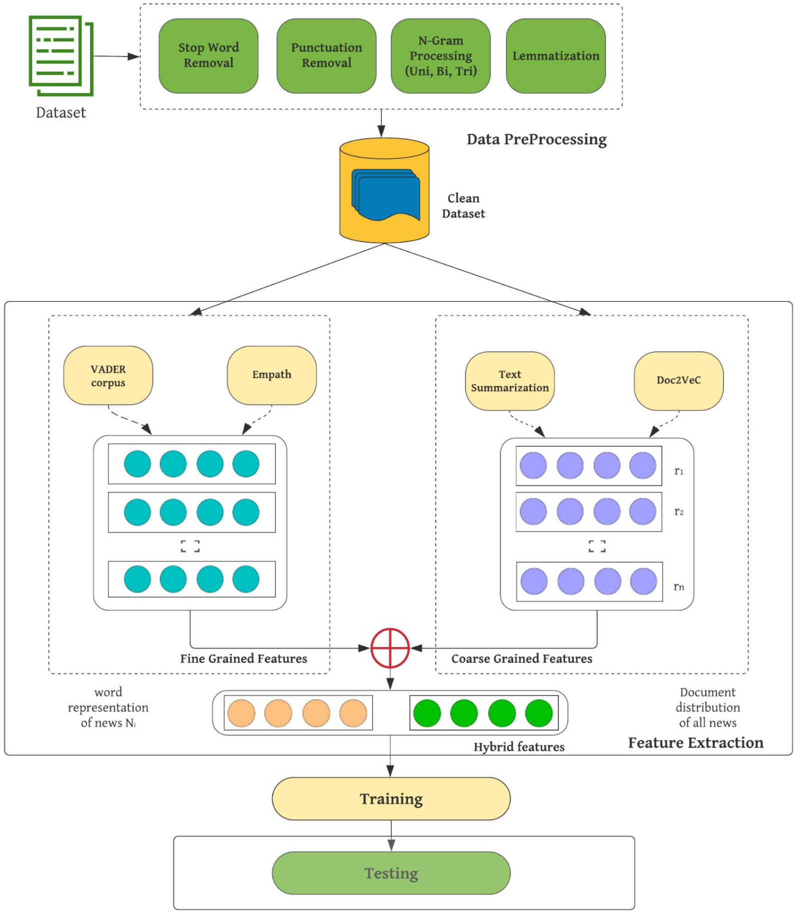
Literature Review

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

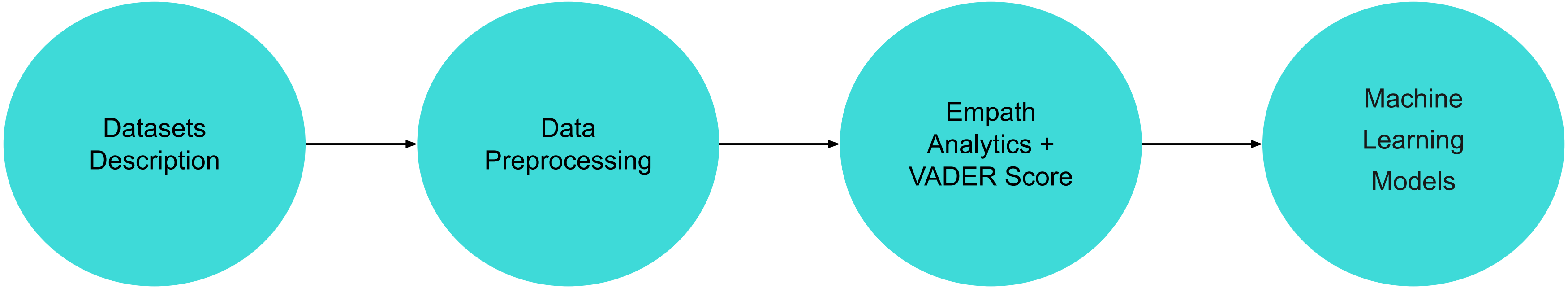
Fine and Coarse Grain Features

- 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)



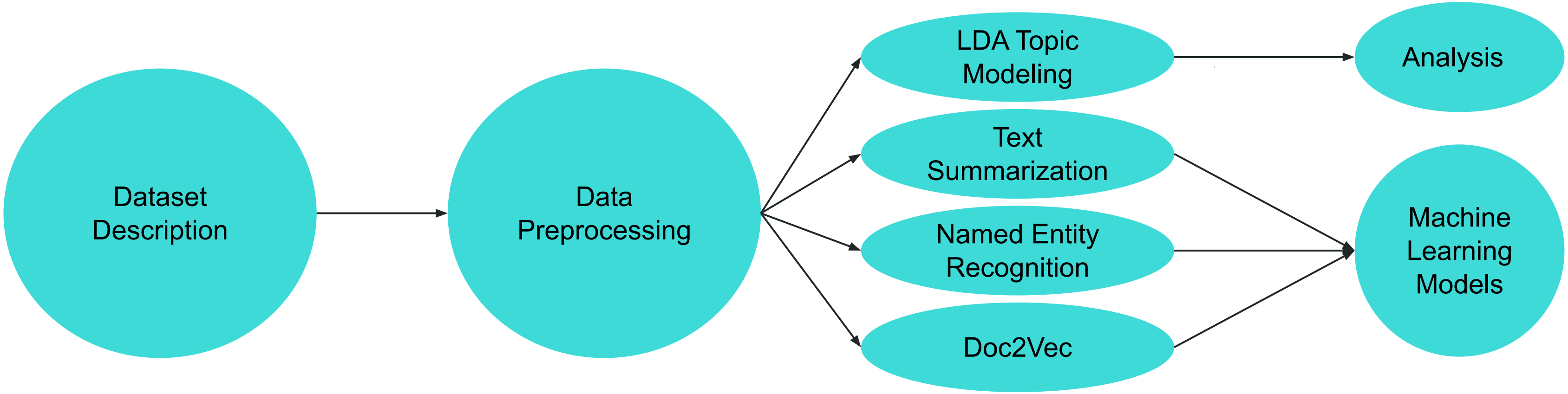
- The features of the dataset are title, text, subject, date, category.
  .

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

- Tool for analyzing text across lexical categories.
- Classifies into around 200 attributes.
- sentiment **score(VADER)**

- Train models on various dataset discussed further.

Solution Flow (Coarse Grained)



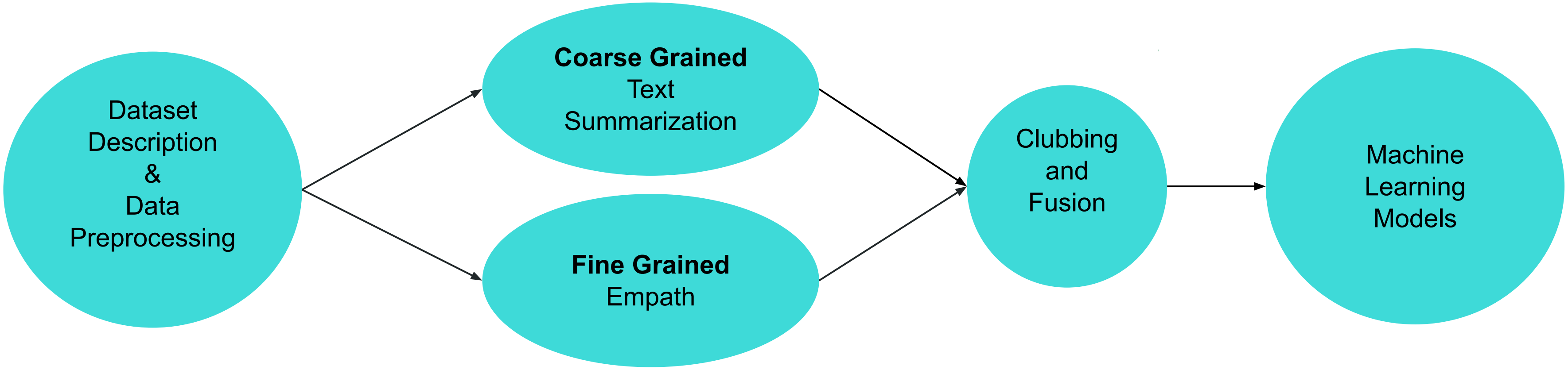
- The features of the dataset are title, text, subject, date, category.
.

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

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

- Train models on various dataset discussed further.

Solution Flow (Fusion)



- The features of the dataset are title, text, subject, date, category.
- .Lowercasing, Lemmatization, Stop-word removal, etc

- Text Summarization with around 20 topics (CG) and Empath with around 195 features (FG).

- The features from fine grain and coarse grain are mixed.

- Train models on various dataset discussed further.

Algorithm Overview

- **Fine Grained Features**
    - Empath Analytics
    - Vader Score
- **Coarse Grained Features**
    - LDA Topic Modeling
    - Text Summarization(Number Of Topic = [10, 15, 20, 25, 30])
    - Named Entity Recognition(NER)
    - Doc2Vec
- **Fusion**
    - Text Summarization(Number Of Topics = [20]) + Empath Analytics

Dataset Analysis

Sr. No.	Dataset	Real	Fake	Total
1	Kaggle	4000	4000	8000
2	COVID19FN	1230	1591	2821
3	Politifact	374	514	888

- The experimentation is carried out on three standard publicly available datasets:
  - Kaggle News Dataset
  - COVID19FN
  - PolitiFact

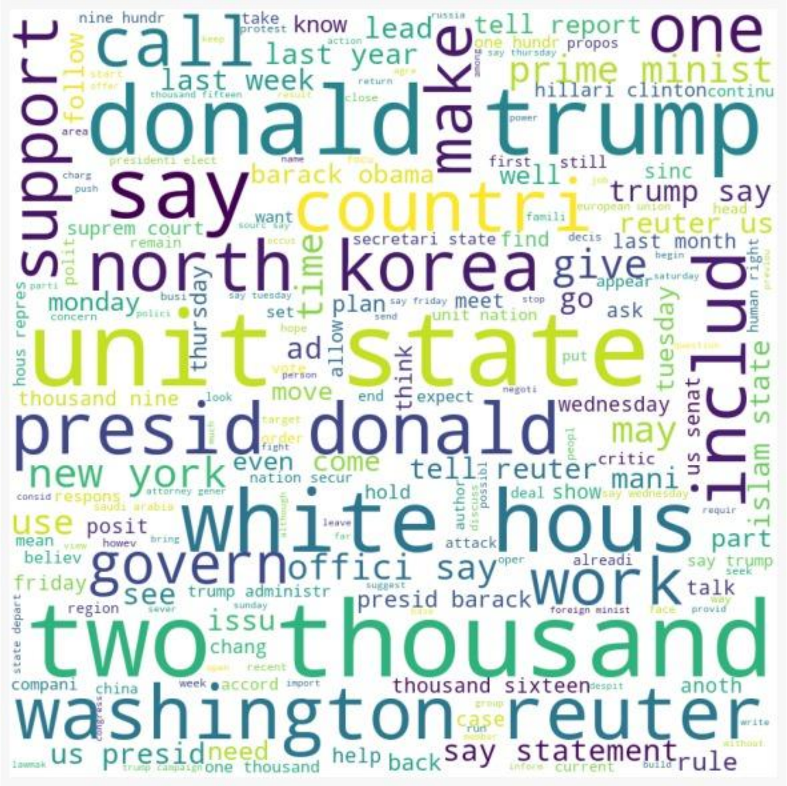
Explainability

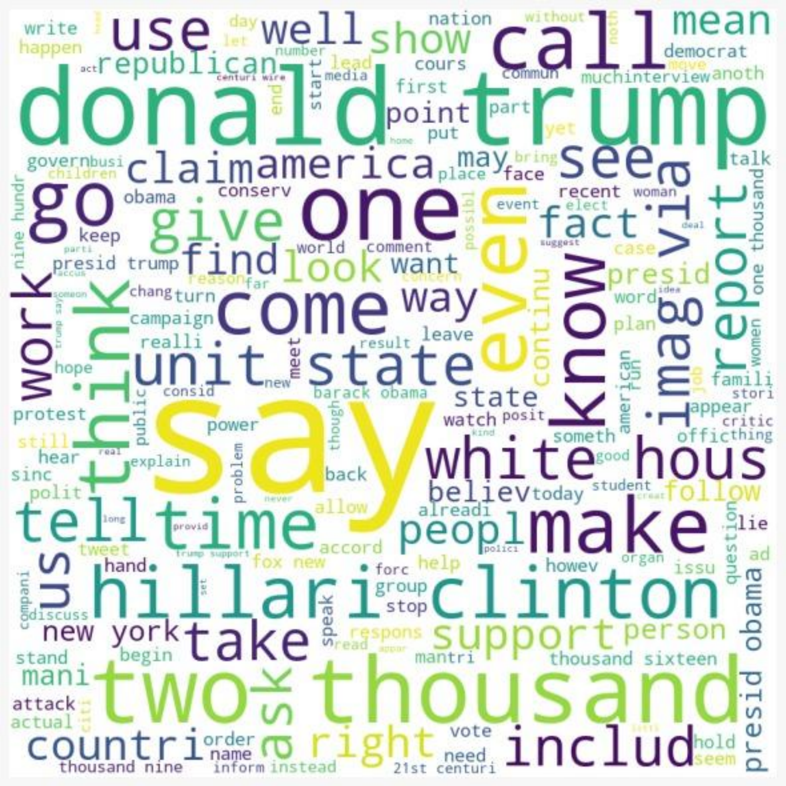
- Technique used to extract which features in the data are most important, how much does each feature effect the prediction.
- A single column of the validation data is randomly shuffled, leaving the target and all other columns in place, and the accuracy of predictions is then checked.
- A column on which model relied heavily for predictions is shuffled then accuracy suffers quite a lot.

Explainability (contd.)

Weight	Feature
0.0893 ± 0.0052	['think', 'want', 'right', 'tweet', 'america']
0.0228 ± 0.0003	['post', 'facebook', 'covid', 'claim', 'social']
0.0225 ± 0.0042	messaging
0.0170 ± 0.0025	speaking
0.0137 ± 0.0025	['government', 'administration', 'unite', 'fund', 'company']
0.0054 ± 0.0028	['north', 'korea', 'trade', 'south', 'unite']
0.0045 ± 0.0011	['clinton', 'hillary', 'campaign', 'election', 'vote']
0.0039 ± 0.0006	['senate', 'vote', 'republican', 'republicans', 'democrats']
0.0025 ± 0.0012	swearing_terms
0.0020 ± 0.0016	giving
0.0020 ± 0.0007	['obama', 'barack', 'administration', 'years', 'claim']
0.0019 ± 0.0013	ridicule
0.0018 ± 0.0008	worship
0.0017 ± 0.0014	leader
0.0016 ± 0.0010	['video', 'police', 'claim', 'share', 'man']
0.0016 ± 0.0009	morning
0.0015 ± 0.0006	eating
0.0015 ± 0.0007	hate
0.0014 ± 0.0015	healing
0.0014 ± 0.0009	clothing

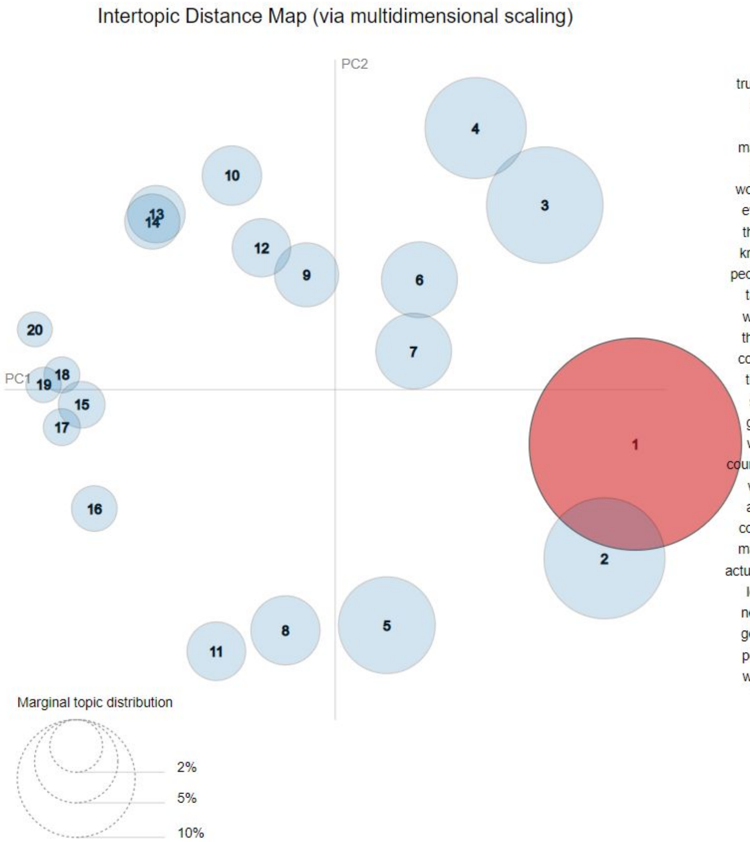
Analysis Of Results (Fine Grain)

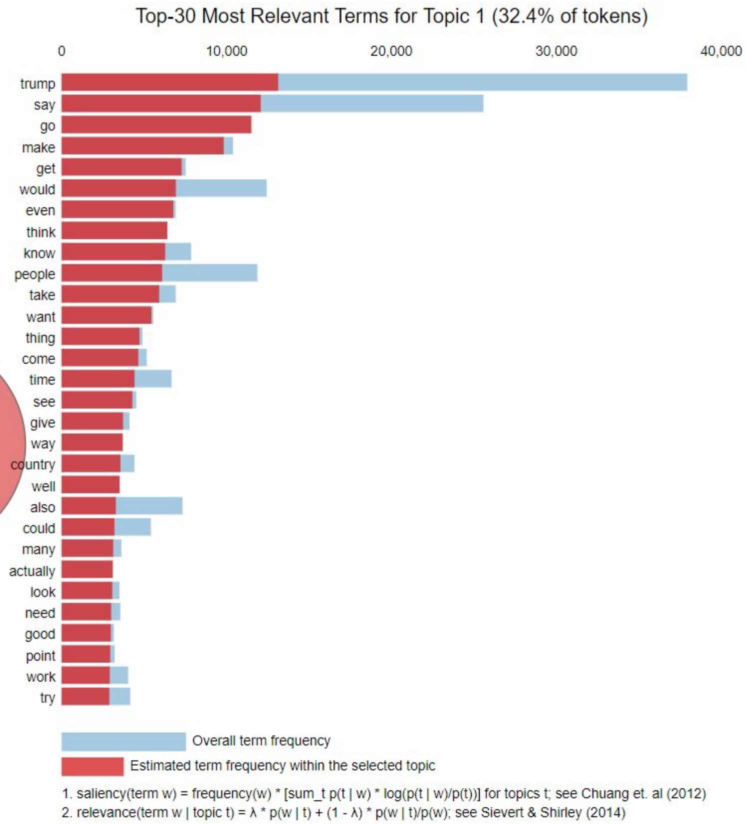




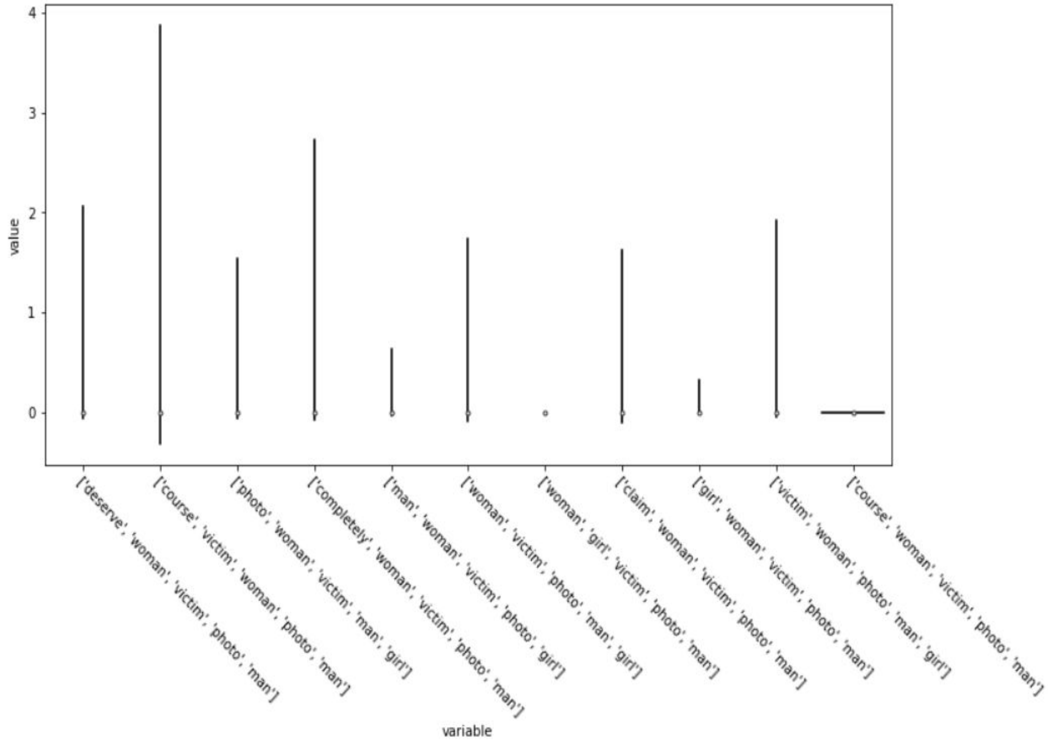
True News Word Cloud **False News Word Cloud**

Analysis Of Results (Coarse Grain - LDA)





Analysis Of Results (Coarse Grain - TS)

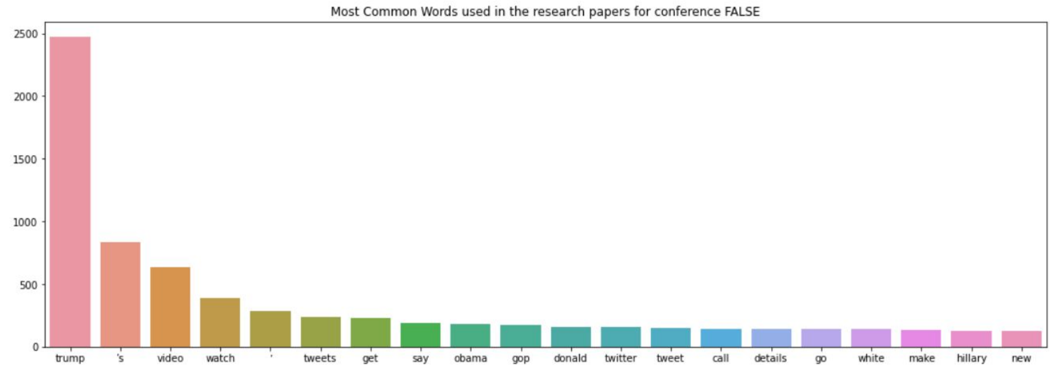


Text Summarization Topics

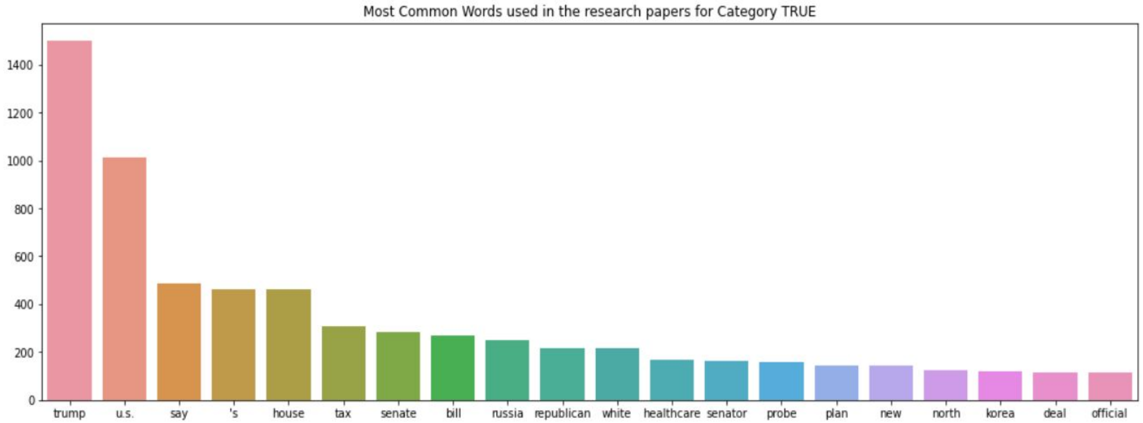
```
{0: ['women', 'know', 'right', 'don', 'going'],

 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 Words



NER True Words

Analysis Of Results (Coarse Grain - NER)

While Donald Trump PERSON was on the campaign trail he blasted the Bush PERSON administration, saying they lied about the existence of weapons of mass destruction, a talking point which led our country to invade a sovereign nation. Trump PERSON has blasted politicians who voted for the Iraq War EVENT . Trump PERSON argued that the move to topple Saddam Hussein PERSON may have been the worst decision in presidential history. But now, he s bringing Dick Cheney PERSON on board to help ensure that Rex Tillerson PERSON is confirmed next year DATE as Trump PRODUCT s secretary of state.Some Republicans NORP have made it clear they have reservations about Tillerson PERSON s relationship with Russian NORP President Vladimir Putin PERSON . So, Cheney PERSON is now Trump s PRODUCT surrogate in order to serve as a bridge between the Trump PERSON team and skeptical Republicans NORP . Cheney PERSON is a former oil executive and longtime friend of Tillerson PERSON s, according to Politico PRODUCT .It s a scenario no one could have possibly foreseen: that one of the key architects of the Iraq War EVENT , which Trump PERSON slammed on the campaign trail, is now being enlisted as an emissary for a man Trump PERSON wants to help steer his ship of state.Wrong, Politico PRODUCT . At the present time, nothing is surprising anymore after Trump LOC s election.Another transition aide said Cheney PERSON s imprimatur may serve as a good housekeeping seal of approval with Republican NORP skeptics. And indeed, Florida GPE Sen. Marco Rubio PERSON received a call from the former vice president earlier this week DATE . The goal: To move Marco PERSON the right way, according to a source familiar with the conversation. Rubio ORG will cast a pivotal vote on the Senate Foreign Relations Committee ORG , which must approve the nomination before it proceeds to the full Senate ORG . Cheney PERSON is also in close contact with senior Trump PRODUCT aides and speaks frequently with Mike Pence PERSON who has said he hopes to model his vice presidency on Cheney PERSON s. Mike PERSON relishes the advice, said a senior transition aide. The aide added that Cheney PERSON wants to be helpful to Trump PRODUCT s administration.You can t make this stuff up. Please just wake me up in four years DATE . That swamp is now filled with terrifying creatures.Photo by Chip Somodevilla PERSON via Getty ORG

Simulation And Results

Model Type Algorithm ML-Models			Kaggle Dataset		Covid Dataset		PolitiFact	
			Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Coarse-Grained	Doc2Vec	Random Forest	0.5733	0.5159	0.7905	0.7906	0.7078	0.6804
		Logistic Regression	***0.9160***	***0.9154***	0.8047	0.8051	0.8127	0.8137
		Gradient Boosting	0.7926	0.7865	***0.8059***	***0.8006***	***0.7790***	***0.7805***
	Text Summarization	Topic = 10	KNN-3	0.8903	0.89	0.9004	0.91	0.7865	0.79
			Random Forest	**0.9056**	**0.91**	0.9218	0.92	***0.8464***	***0.85***
			Gradient Boosting	0.905	0.90	0.9194	0.92	*0.8464*	*0.85*
		Topic = 15	KNN-3	0.893	0.891	0.911	0.91	0.8089	0.81
			Random Forest	0.920	0.921	0.923	0.92	0.8576	0.86
			Gradient Boosting	***0.923***	***0.922***	***0.928***	***0.93***	***0.8614***	***0.86***
		Topic = 20	KNN-3	*0.8833*	*0.88*	*0.9102*	*0.91*	0.8352	0.84
			Random Forest	***0.9203***	***0.92***	***0.9445***	***0.94***	**0.8726**	**0.87**
			Gradient Boosting	*0.913*	*0.91*	*0.9397*	*0.94*	0.8576	0.86
		Topic = 25	KNN-3	*0.8563*	*0.86*	*0.8902*	*0.89*	0.8089	0.81
			Random Forest	*0.9193*	*0.92*	***0.9327***	***0.93***	***0.8614***	***0.86***
			Gradient Boosting	***0.9203***	***0.92***	*0.9327*	*0.93*	0.8389	0.84
		Topic = 30	KNN-3	*0.8593*	*0.86*	*0.8795*	*0.88*	0.7827	0.78
			Random Forest	***0.9176***	***0.92***	*0.9397*	*0.94*	0.8127	0.81
			Gradient Boosting	*0.9167*	*0.92*	***0.9421***	***0.94***	***0.8614***	***0.86***
	NER	Gradient Boosting	0.890	0.892	0.936	0.94	0.8775	0.88
		Random Forest	0.931	0.930	***0.952***	***0.95***	***0.9149***	***0.91***
		Linear SVM	***0.949***	***0.951***	0.947	0.95	0.8673	0.87

Coarse Grained Results

Simulation And Results

Fine-Grained	Empath Analytics	LDA	0.901	0.901	0.889	0.89	0.8277	0.83
		Random Forest	0.908	0.910	***0.923***	***0.92***	*0.8127*	*0.81*
		Gradient Boosting	***0.928***	***0.931***	0.921	0.92	**0.8726**	**0.87**
	Empath + VADER	LDA	0.9073	0.91	0.888	0.89	0.8127	0.81
		Random Forest	0.9086	0.91	***0.925***	***0.93***	*0.8314*	*0.83*
		Gradient Boosting	***0.9326***	***0.93***	0.921	0.92	**0.8764**	**0.88**
Coarse and Fine-grained fusion	TextSummarization + Empath	LDA	0.884	0.880	0.891	0.89	0.7902	0.79
		Random Forest	0.895	0.900	0.926	0.93	***0.8389***	***0.84***
		Gradient Boosting	***0.896***	***0.900***	***0.926***	***0.93***	0.8352	0.84

Fine Grained Result

Fusion Result

Model Type Algorithm ML-Models			**All Merged**	
			Accuracy	**F1-Score**
Coarse-Grained	Text Summarization Topic = 20	KNN-3	0.8655	0.87
		Random Forest	**0.8842**	**0.88**
		Gradient Boosting	0.8751	0.88
Fine-Grained	Empath Analytics	LDA	0.8340	0.83
		Random Forest	*0.8699*	*0.87*
		Gradient Boosting	**0.8732**	**0.87**
Coarse and Fine-grained fusion	TextSummarization + Empath	LDA	0.8881	0.89
		Random Forest	0.8988	0.9
		Gradient Boosting	***0.9056***	***0.91***

all results on merged dataset

Baseline Comparison

Dataset	Author	Author's Approach	Paper's Accuracy	Our Approach	Our Accuracy
Fake and Real News Dataset(Kaggle)	Ahmed, H. et. al.	n-gram features and the LSVM algorithm	87.0	Empath + Text Summarization fusion	**89.6**
Liar Dataset	Wang, W. Y. et. al.	SVM, Bi-LSTMs	26.1	Text Summarization + SVM	**55.6**

- We also list down the efficiency and approach we executed to attain the same.
- Fake and Real News Dataset (Kaggle) and Liar Dataset are used for the basis of comparison with referenced papers.

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.
- Comprehensive experiments were designed and implemented based on three existing standard datasets.
- **Link to Report :-** Click here

Future Works

INTEGRATE	Integrate with appropriate platforms ie. chrome extension
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.
[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.

References

[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.
[30] Fuller, C. M., Biros, D. P., Wilson, R. L. (2009). "Decision support for determining veracity via linguistic-based cues. Decision Support Systems", 46(3), 695-703.
[31] Sriram, S. (2020). "An Evaluation of Text Representation Techniques for Fake News Detection Using: TF-IDF, Word Embeddings, Sentence Embeddings with Linear Support Vector Machine".
[32] Le, Q., Mikolov, T. (2014, June). "Distributed representations of sentences and documents" In International conference on machine learning (pp. 1188-1196). PMLR.
[33] Fake and Real News Dataset(Kaggle) Ahmed, H., Traore, I., Saad, S. (2018). Detecting opinion spam and fake news using text classification. Security and Privacy, 1(1), e9.
[34] Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.
[35] Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).


Thank
you!!