NLP Literature Survey:

Charvi:

Sl. No	Link	Title	Authors	Aims	Dataset	Method	Challenges/Gaps	Accuracy
1.	https://link.springer.com/article/10.1007/s11257-016-9171-0	Computational personality recognition in social media	Golnoosh Farnad et. al	Classify Personality traits like extroversion, openness, agreeableness ,Conscientiousness, Emotional stability	Uses data from the myPersonality project. Twitter, Facebook and Youtube.	It uses regression methods on after various features have been extracted. LIWC, NRC, MRC and other psycho-lingual feature extraction techniques.	Very few personality traits are predicted. Simple regression model.	RMSE 0.807 ishh
2.	https://arxiv.org/abs/1901.09672	Personalized Dialogue Generation with Diversified Traits	Yinhe Zheng et. al	To predict gender, age and location as a key-value pair using character traits present in dialogues	Unique generated dataset called PersonalDialog	Seq2Seq models along with fusion with personality attention	Predicts only Age, Gender, Location	96%
3.	https://journalofbigdata.springeropen.com/articl	Text based personality prediction from	Hans Christian et al.	Detailed Big Five personality detection	Twitter and Facebook data	Essemble on RoBERTa+BERT+XLNet	-	86%

	es/10.1186/s40537-021-00459-1	multiple social media data sources using pre-trained language model and model averaging						
4.	https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8909523	Personality Recognition in Conversations using Capsule Neural Networks	Esteban A. Ríssola et. al	Big Five personality traits highlighting	Conversational dataset- not publicly available	Capsule Neural Networks using custom routing algorithm	Categorical prediction rather than continuous.	83% Recall

Chaitra:

Sl. No	Link	Title	Authors	Aims	Dataset	Method	Challenges/Gaps	Accuracy
1.	https://dl.acm.org/doi/abs/10.1145/3055601.3055603	A Machine Learning Approach to Demographic Prediction using Geohashes	Avipsa Roy et. al	Develop a model which will predict the age and gender of users in separate groups with respect to their most frequently visited location.	Real dataset of mobile phone users collected and shared by a telecommunication provider	Discriminant Analysis approach from the Python scikit-learn, FPGrowth model, Decision Trees	Demographic information such as age and gender are mostly unavailable to app developers for open access due to privacy concerns	71.62% and 96.75% for predicting gender and age groups of the users respectively.
2.	https://www	Personality	Alam Sher Khan	This work	Publically	KNN, Decision	The skewness of the	XGBoost

	.researchgate.net/publication/340399695_Personality_Classification_from_Online_Text_using_Machine_Learning_Approach	Classification from Online Text using Machine Learning Approach	et.al	provides the basis for developing a personality identification system which could assist organizations for recruiting and selecting appropriate personnel.	available benchmark dataset from Kaggle	Tree, Random Forest, MLP, Logistic Regression (LR), SVM, XGBoost, MNB and Stochastic Gradient Descent (SGD).	dataset is the main issue.	classifier is outstanding by achieving more than 99% precision and accuracy.
3.	https://dl.acm.org/doi/abs/10.1145/2663204.2663272	Statistical Analysis of Personality and Identity in Chats Using a Keylogging Platform	Giorgio Roffo et. al	semi-structured chats between 50 subjects, whose personality traits have been analyzed through psychometric tests, and a single operator, for a total of 16 hours of conversation.	The data collection has been based on a public social network, where the text chatting interface has been equipped with keylogging functionalities.	LexicalFeatures ,SyntacticFeatures,Turn-takingFeatures, Regression, Support Vector Regressor (-SVR) with RBF kernel.	-	63.5% nAUC
4.	https://www.sciencedirect.com/science/article/pii/S2451958820300166	The demographics of computer-mediated communication: A review of social media demographic	Sarah Gambo et. al	SNSs(Social Network Sites) usage distributed amongst various demographic groups, gender and SNS usage,	Facebook, Whatsapp, Instagram analysis	Social Networking Site , Percentage of internet users using it Demographic group this service is most	Tracking the global digital gender gap	This review study explored the demographics of SNS usage using three case-studie

		trends among social networking site giants		and SNS and age factor.		appealing to.		s: Facebook, Instagram, and WhatsApp

Shruthi:

Sl. No	Link	Title	Authors	Aims	Dataset	Methods	Challenges/Gaps	Accuracy
1.	https://ieeexplore.ieee.org/abstract/document/9121971	Predicting Personality Using Answers to Open-Ended Interview Questions	Madhura Jayaratne; Buddhi Jayatilleke	Aims to show that textual content of answers to standard interview questions related to past behaviour and situational judgement can be used to reliably infer personality traits	Data from over 46,000 job applicants who completed an online chat interview that also included a personality questionnaire based on the six-factor HEXACO personality model to self-rate their personality	HEXACO model of personality as the underlying personality representation model, open-voc	Used only the semantic level features (terms, topics etc).Exploring whether other types of features, such as the use of parts of speech (POS), readability, formality, use of emojis etc. can further increase the accuracy. Testing the performance of other available regression algorithms, including neural network approaches, may help increase the accuracy of the regression models	Terms and topics based text representation achieved the best accuracy, an average correlation of 0.387 over other representation methods.

						abulary approach in nlp, regression model		
2.	https://arxiv.org/abs/2004.04460	PANDORA Talks: Personality and Demographics on Reddit	Matej Gjurkovic Mladen Karan Iva Vukojevi ´ c Mihaela Bo ´ snjak Jan ˇ Snajder	To present the dataset PANDORA with both personality and demographic labels. To showcase the usefulness of this dataset on three experiments. Present benchmark prediction models for all personality and demographic variables.	Reddit comments of 10k users partially labeled with three per sonality models and demographic s (age, gender, and location), including 1.6k users labeled with the well established Big 5	They showcased the usefulness of PANDORA with three experiments, showing (1) how more readily available MBTI/Enneagram labels can be used to estimate Big 5 traits, (2) that a gender classifier trained	The poor performance of deep learning baseline models, the rich set of labels, and the large number of comments per user in PANDORA suggest that further efforts should be directed toward efficient user representations and more advanced deep learning architectures.	-

					personality model	on Reddit exhibits bias on users of certain personality traits, and (3) that certain psycho-demographic variables are good predictors of propensity for philosophy of Reddit users.		
3.	https://www.emerald.com/insight/content/doi/10.1108/ACI-03-2021-0054/full/html	Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia	Ema Utami, Irwan Oyong, Suwanto Raharjo, Anggit Dwi Hartanto, Sumarni Adi	Analyzing profile data from personal social media accounts reduces data collection time, as this method does not require users to fill any questionnaires. A pure natural language processing (NLP) approach can give decent results, and its reliability can be improved by combining it with machine learning	Raw data were derived from a predefined dominance, influence, stability and conscientious (DISC) quiz website, returning 316,967 tweets from 1,244 Twitter accounts “filtered to include only personal and Indonesian-language accounts”	Data collection, pre-processing, feature preparation using nlp techniques. Model fitting using default parameter and initial data distribution. Data resampling using Random Undersampling, SMOTE, and SMOTETomek. Hyper-parameter tuning on SVC, Random Forest Classifier,	Observation analysis of other social media platforms (such as Facebook, Instagram and LinkedIn) is also commonly practiced during the employee selection process which they havent considered in the paper.A combination of textual and visual information derived from those platforms might be able to provide more comprehensive and better classification results	Among resampling techniques, SMOTETomek returning the best performance. Hyperparameter-tuned support vector classifier outperformed several supervised and ensemble learning algorithms, with an

						Gradient Boosting Classifier, AdaBoost Classifier, and Decision Tree Classifier and best model evaluation was performed.		F1-score of 56.43%.
4.	https://ieeexplore.ieee.org/abstract/document/9760970?casa_token=3KODThPj44AAAAAA:9__0V7mCyJlKV_oAvsMkz151uBvCFmPbUlK1Ho8vzWRw3hBRzNGQSjicPsE3bgPx9465nceDyxqh	Hybrid Machine Learning Technique for Personality Classification from Online Text using HEXACO Model	P. William; Abhishek Badholia; Brijesh Patel; Manoj Nigam	To find a suitable machine learning technique for decision assistance for personality classification that delivers accurate and comprehensible results while remaining within budget constraints	Publicly accessible benchmark dataset from Kaggle	Data Level Re-Sampling was done to balance the class instances by rescaling the training datasets. Random Forests and Stacking ensemble methods for training. K Fold Cross Validation for testing	For a more accurate evaluation of the characteristics, the personality models should be compared to the other models	The findings indicate that for I/E and S/N characteristics, we achieved more than 99 percent precision. and accuracy, and for T/F and J/P dimensions, we received approximately 95 percent accuracy. The KNN classifier, on

								the other hand, had a lower overall performanc e.

Suraj:

Sl. No	Link	Title	Authors	Aims	Dataset	Method	Challenges/Gaps	Accuracy
1.	https://arxiv.org/pdf/2210.07871.pdf	One Graph to Rule them All: Using NLP and Graph Neural Networks to analyse Tolkien's Legendarium	Vincenzo Perri, Lisi Qarkaxhija, Albin Zehe, Andreas Hotho, Ingo Scholtes	Analyse micro and macro level details of characters within the corpus.	J.R.R. Tolkien's Legendarium	Graph neural networks for establishing connections and NLP in order to understand personality.	Embedding and co-occurrence prediction. Predicting missing links in a graph	The GCN model is able to accurately predict character classes, reaching an f1-score of ≈ 79.7%, a precision of ≈ 78.6% and a recall of ≈ 82.6%
2.	https://digitalcommons.bard.edu/cgi/viewcontent.cgi?article=1097&conte	Heroes, Villains, and the In-Between: A Natural Language Processing Appr ocessing	Ruby Alling Ostrow	Aim is to determine if it is possible to develop a computational model to extract	Dataset of fifty Grimm's fairy tales	SpaCy's dependency parser. Adolfo & Ong 2019. Stanford	Each tale has different number of characters making it hard to determine whether the identified main character is correct.	On 13 average, there was a precision of 0.88, a recall of

	xt=senproj_s2022	Approach to Fairy Tales		the defining features from the odd and unpredictable nature of this classic form of literature.		CoreNLP library	Adjective recognition is tricky, given the limitation to copular verbs, given the lack of knowledge of where an adjective describing a character may appear in the sentence.	0.91, and a F1 of 0.89
3.	https://www.mdpi.com/2414-4088/4/1/9	Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator	Mohammad Hossein Amirhosseini, Hassan Kazemian	Aim is to develop a new machine learning method for personality type prediction based on the MBTI.	Myers–Briggs Personality Type Dataset from Kaggle	The natural language processing toolkit (NLTK) and XGBoost which is an optimized distributed Gradient Boosting library in Python. Numpy, XGBoost and sklearn were used to create the Gradient Boosting Model	-	Introversion (I)–Extroversion (E) accuracy = 78.17% Intuition (I)–Sensing (S) accuracy = 86.06% Feeling (F)–Thinking (T) accuracy = 71.78% Judging (J)–Perceiving (P) accuracy = 65.70%
4.	https://ieeexplore.ieee.or	Analysis of Personality Traits	Tejas Pradhan; Rashi Bhansali;	This paper focuses on	A labeled dataset with	Naïve Bayes, SVM,	There is a bias in the model as some	Naïve Bayes

	g/abstract/document/9183090?casa_token=dEfiZgxIHWMAAAAA:HPty4KvXIo9aaL7YYuzDvwf-mqXzyrwBOxNXJXpV8GMzd9ajVrqJOvgzbLbMM_Bd1W14N03MFhJU	using Natural Language Processing and Deep Learning	Dimple Chandnani; Aditya Pangaonkar	automating personality testing and analysis with the help of Neural Networks by using images instead of questions.	user responses on social media along with their personality type is used for analysis	CNN	personality types are more common than the others. This can be tackled by scraping more data of the minority personality types. Additional models like recurrent neural networks can be used to improve the prediction accuracy by taking into consideration, the past results. The website becomes slow when there is too much traffic because the neural network model makes the back end heavy.	Accuracy = 32.6% SVM Accuracy = 57.9% CNN accuracy = 81.4%