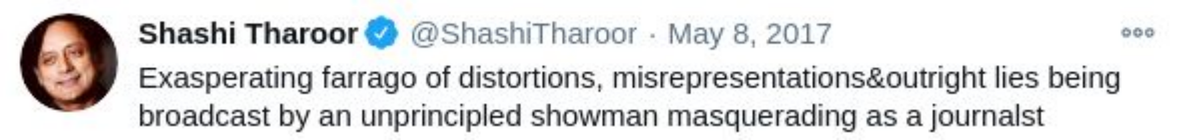
WOMEN WHO CODE

CONNECT Forward

Lexical Simplification of Hindi Text

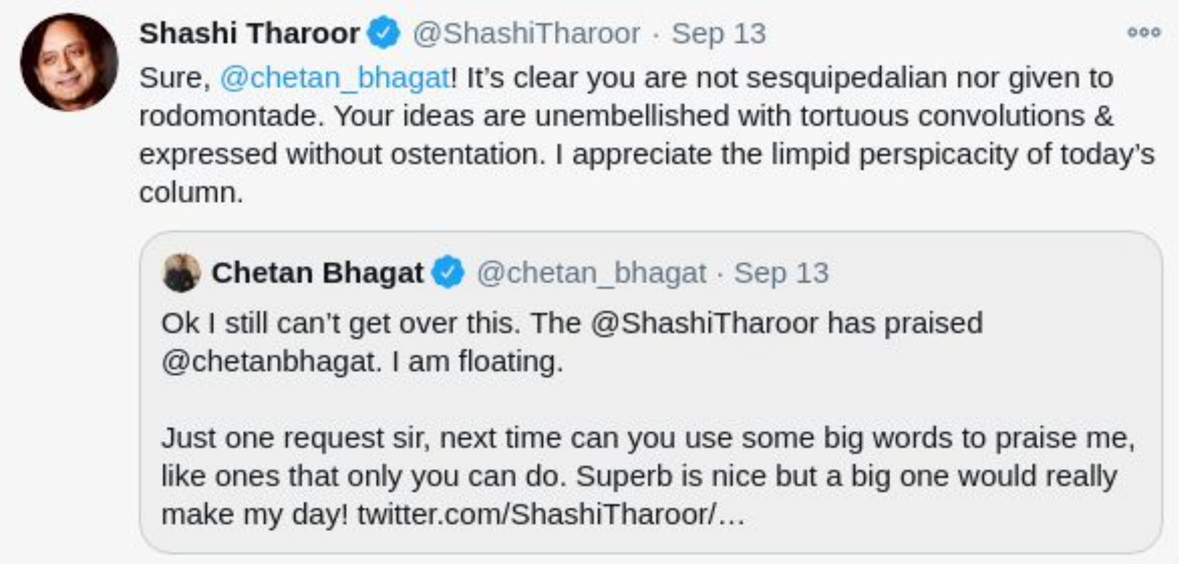
Gayatri Venugopal
Twitter: **@speakingg3**
https://gayatrivenugopal.wordpress.com
gayatrivenugopal3@gmail.com
https://github.com/gayatrivenugopal/wwcode-connectforward

Text Simplification...Why?



Shashi Tharoor @ShashiTharoor · May 8, 2017

Exasperating farrago of distortions, misrepresentations&outright lies being broadcast by an unprincipled showman masquerading as a journalst



Shashi Tharoor @ShashiTharoor · Sep 13

Sure, @chetan_bhagat! It's clear you are not sesquipedalian nor given to rodomontade. Your ideas are unembellished with tortuous convolutions & expressed without ostentation. I appreciate the limpid perspicacity of today's column.

Chetan Bhagat @chetan_bhagat · Sep 13

Ok I still can't get over this. The @ShashiTharoor has praised @chetanbhagat. I am floating.

Just one request sir, next time can you use some big words to praise me, like ones that only you can do. Superb is nice but a big one would really make my day! twitter.com/ShashiTharoor/...

Sources

https://twitter.com/shashitharoor/status/1305097520981225472?lang=en

https://twitter.com/shashitharoor/status/861608665517895680?lang=en

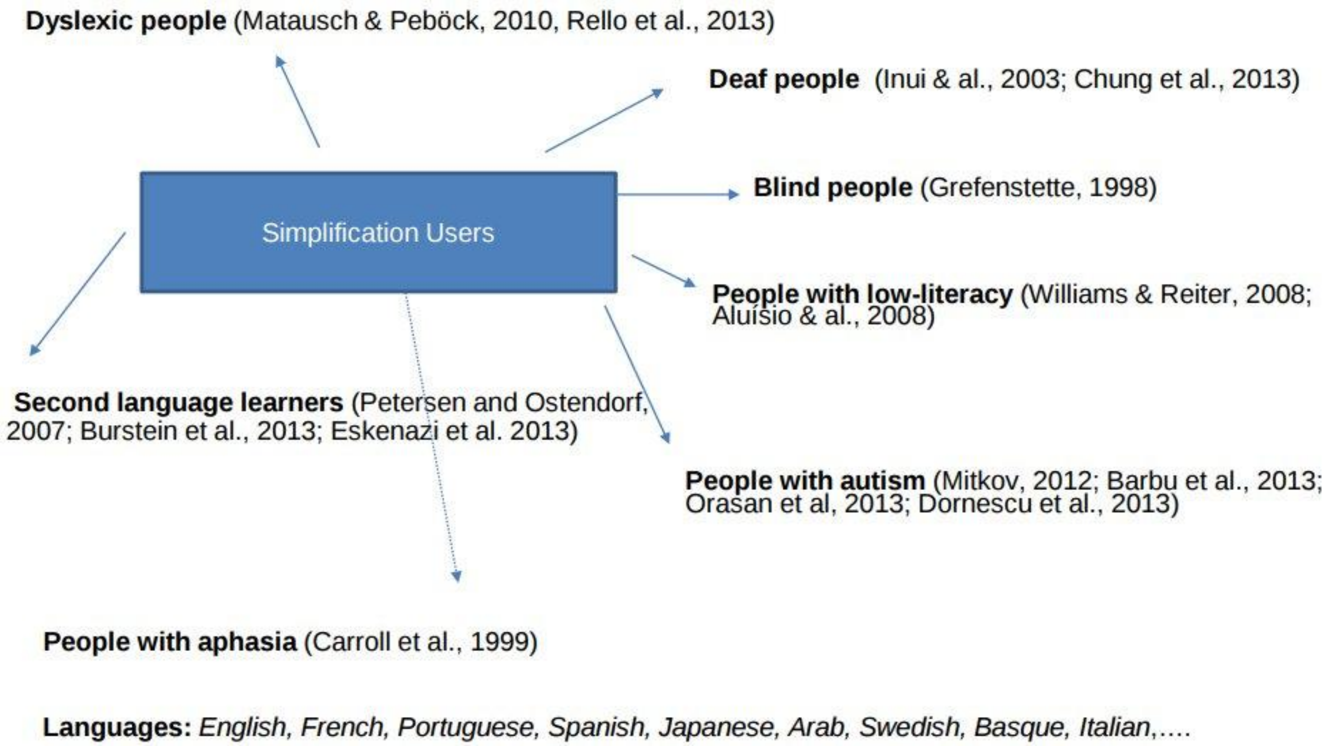
Nomination Acceptance Speeches

Tia Dufour / White House • Gage Skidmore / CC BY-SA 2.0

Flesch Reading Ease (100% = simple)	60%	72.3%
Length	6944 words	3196 words

Sources
https://www.bloombergquint.com/gadfly/biden-and-trump-nomination-speeches-are-two-clashing-word-clouds

Text Simplification and its Applications



Simplification Users

Dyslexic people (Matausch & Pebӧck, 2010, Rello et al., 2013)

Deaf people (Inui & al., 2003; Chung et al., 2013)

Blind people (Grefenstette, 1998)

People with low-literacy (Williams & Reiter, 2008; Aluísio & al., 2008)

People with autism (Mitkov, 2012; Barbu et al., 2013; Orasan et al, 2013; Dornescu et al., 2013)

Second language learners (Petersen and Ostendorf, 2007; Burstein et al., 2013; Eskenazi et al. 2013)

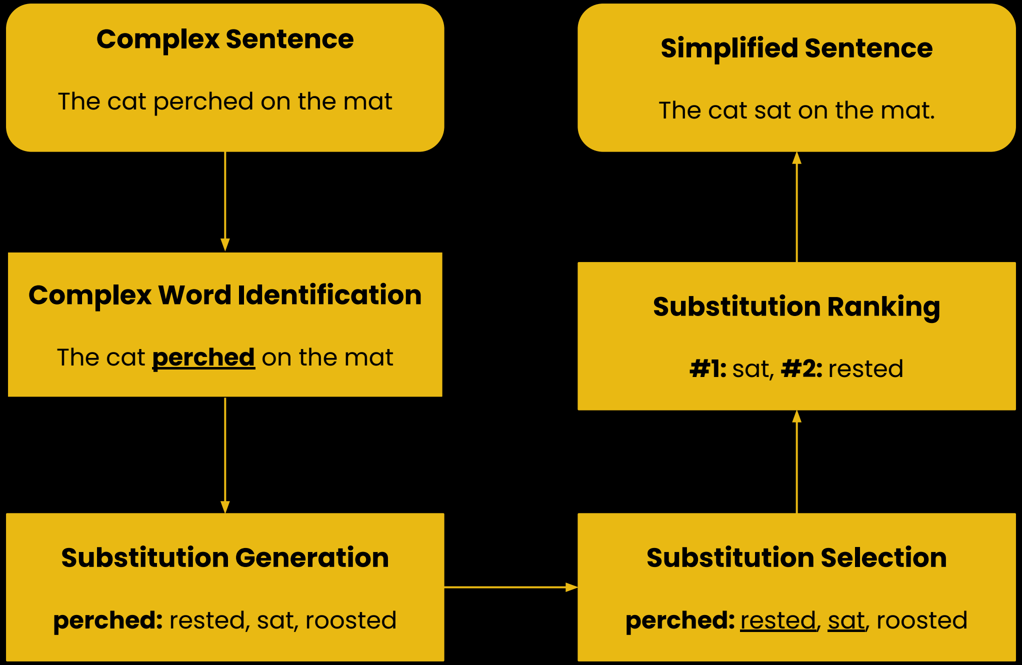
People with aphasia (Carroll et al., 1999)

Languages: *English, French, Portuguese, Spanish, Japanese, Arab, Swedish, Basque, Italian,….*

Sources

https://zenodo.org/record/2593329#.XIzyaayYW02 presented at the Department of Information and Communication Technologies, Universitat Pompeu Fabra, in the context of the María de Maeztu data-driven knowledge extraction strategic research program (MDM-2015-0502), by Horacio Saggion and Sanja Štajner in 2019.

Lexical Simplification Pipeline



Complex Sentence

The cat perched on the mat

Complex Word Identification

The cat **perched** on the mat

Substitution Generation

perched: rested, sat, roosted

Substitution Selection

perched: rested, sat, roosted

Substitution Ranking

#1: sat, **#2:** rested

Simplified Sentence

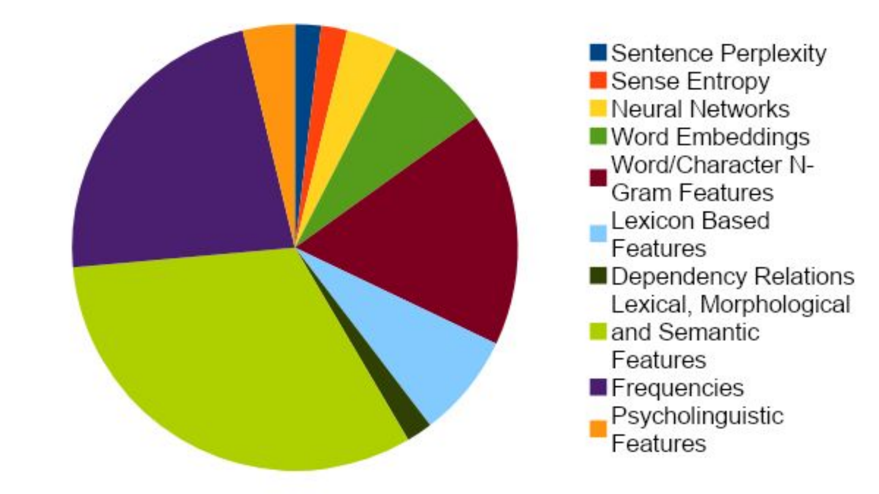
The cat sat on the mat.

Sources

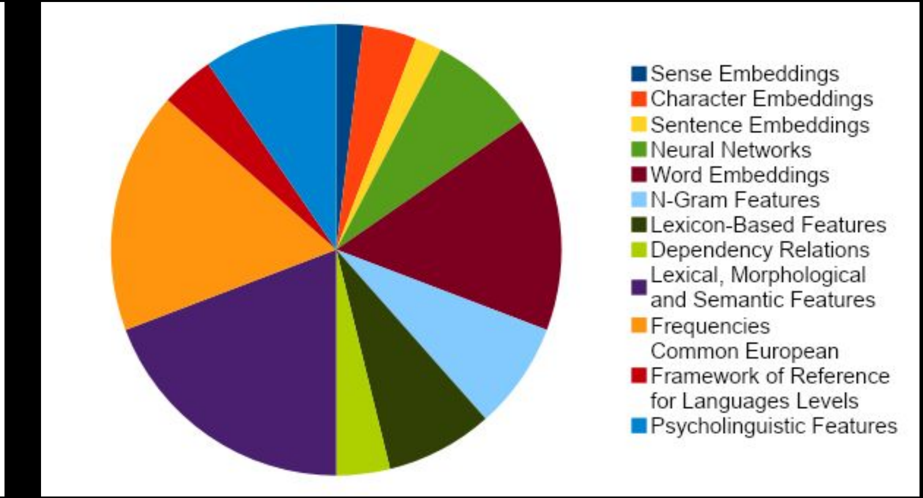
Paetzold, G. (2015, June). Reliable lexical simplification for non-native speakers. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (pp. 9-16).

Adaptation of the pipeline proposed by Shardlow, M. (2014, May). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In LREC (pp. 1583-1590).

SemEval Shared Tasks



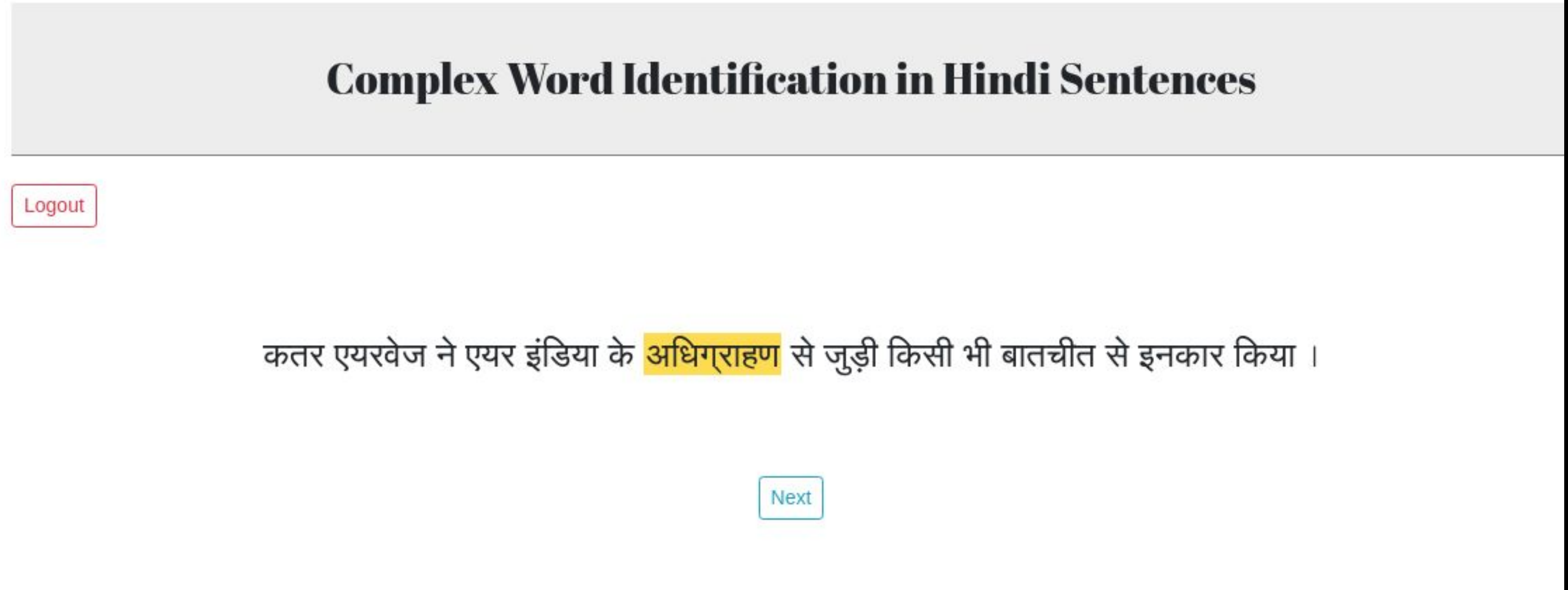
Summary of features employed in CWI Shared Tasks 2016



Summary of features employed in CWI Shared Tasks 2018

Sources
Paetzold, G., & Specia, L. (2016, June). Semeval 2016 task 11: Complex word identification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 560-569); Muhie Yimam, S., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., ... & Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. arXiv, arXiv-1804.

Task 1 - Annotation

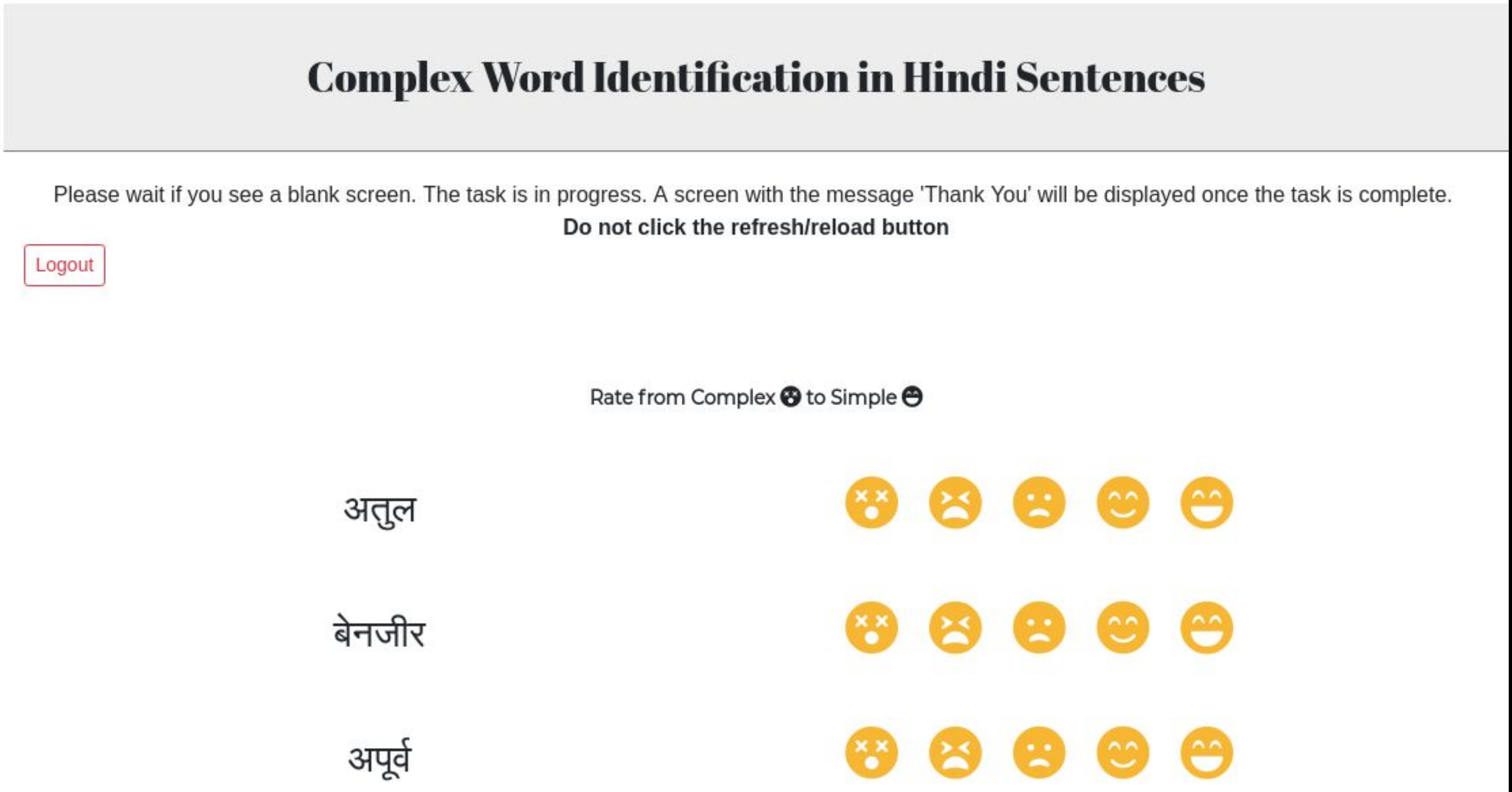


Complex Word Identification in Hindi Sentences

कतर एयरवेज ने एयर इंडिया के अधिग्राहण से जुड़ी किसी भी बातचीत से इनकार किया ।

Screens designed by **Kumar Ashwin (https://krash.dev)**

Task 2 - Rating



Complex Word Identification in Hindi Sentences

Please wait if you see a blank screen. The task is in progress. A screen with the message 'Thank You' will be displayed once the task is complete.

Do not click the refresh/reload button

Logout

Rate from Complex 😵 to Simple 😁

अतुल 😵 😫 😟 😊 😁

बेनजीर 😵 😫 😟 😊 😁

अपूर्व 😵 😫 😟 😊 😁

Screens designed by **Kumar Ashwin (https://krash.dev)**

S.No.	Source	Unique Word Count	Unique Lemma Count	Domain
1	Language Resource#1	145,507	117,309	Aesthetics
2	Language Resource#2	21,335	17,159	Entertainment
3	Language Resource#3	119,313	102,201	Not available
4	Language Resource#4	2,330	1,851	Varied domains
5	Language Resource#5	21,826	18,220	Tourism
6	Language Resource#6	39,351	32,074	Agriculture and Entertainment
7	Language Resource#7	35,018	28,645	Agriculture, Entertainment, Politics and Public, Administration, Sports, Religion, Literature, Aesthetics, Economy
8	Language Resource#8	20,430	16,673	Health
9	Language Resource#9	5,322,602	4,579,200	News

Data Description

Number of words ranked by participants	68,107
Number of unique words ranked by participants	18,186
Number of unique words annotated by at least two participants	12,111
Number of unique words ranked by at least two participants and that are present in our corpus	7,315

Features, Classifiers and Evaluation Metrics

Features	length, number of syllables, frequency of the lemma of the word, number of consonants, number of vowels, number of consonant conjuncts (स◌्कूल -> स + क), number of synsets, number of synonyms, number of hyp
Classifiers	decision tree, support vector classifier, nearest centroid classifier, random forest, extra trees, ada boost, gradient boosting and XG boost
Evaluation Metrics	AUC Scores

- Created a stopword list: https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas
- Normalised feature values
- We used soft voting classification and random search hyperparameter tuning of the models. Receiver Operating Characteristic (ROC) scores were used to tune the models.

Model	AUC Score
Ada	0.776
Tuned Ada	0.781
Extra Trees	0.760
Tuned Extra Trees	0.762
Gradient Boosting	0.783
Tuned Gradient Boosting	0.755
Random Forest	0.770
Tuned Random Forest	0.785
XGBoost	0.785
Tuned XGBoost	0.782
Soft Voting	0.790

Ongoing and Future Work

- Analysis of user data
- Hyperparameter tuning using Grid Search
- Adding fasttext embeddings as a feature for classification
- Treating the task as a token classification task – using data from Task 1
- Word sense disambiguation
- Synonym selection and substitution

Language Resources

Language Resource#1: Aesthetics Corpus

Language Resource#2: The Open Parallel Corpus (n.d.). Retrieved July 12, 2019, from http://opus.nlpl.eu/.

Language Resource#3: CFILT Hindi Corpus (n.d.). Retrieved July 15, 2019, from https://www.cfilt.iitb.ac.in/Downloads.html.

Language Resource#4: Kunchukuttan, A., Mehta, P. & Bhattacharyya, P. (2018). The IIT Bombay English-Hindi Parallel Corpus. Language Resources and Evaluation Conference.

Language Resource#5: English-Hindi Tourism Text Corpus – EILMT (October, 2016). EILMT Consortia, CDAC Pune. Retrieved July 15, 2019, from http://www.tdil-dc.in.

Language Resource#6: Hindi-English Agriculture & Entertainment Text Corpus ILCI-II (May, 2017). ILCI Consortium, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in.

Language Resource#7: Hindi Monolingual Text Corpus ILCI-II (June, 2017). ILCI-II, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in.

Language Resource#8: Hindi-English Health Text Corpus-ILCI (April, 2012). ILCI Consortium, JNU. Retrieved July 15, 2019, from http://www.tdil-dc.in

Language Resource#9: Kunchukuttan, A., Kakwani, D., Golla, S., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. arXiv preprint arXiv:2005.00085.

Gayatri Venugopal
Twitter: **@speakingg3**
https://gayatrivenugopal.wordpress.com
gayatrivenugopal3@gmail.com
https://github.com/gayatrivenugopal/wwcode-connectforward