IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Building a Wikipedia-Based Hindi Fact Verification System

[1]Shaurya Nagar

[1]Student

[1]Genesis Global School, Noida India

Abstract: The proliferation of misinformation across digital platforms in Hindi-speaking regions has placed a dire need for effective and culturally relevant fact-checking systems. To effectively address this growing issue, this paper proposes a novel system to detect and verify claims in Hindi, utilizing public knowledge repositories such as Wikipedia. The system will integrate various Natural Language Processing (NLP) techniques, including Named Entity Recognition (NER), semantic sentence similarity, and Natural Language Inference (NLI). The pipeline utilizes several multilingual models, including mDeBERTa, which have been optimized with Hindi-specific datasets for claim verification. Experimental results demonstrate competitive performance when compared to baseline models, with a focus on scalability and adaptability to the unique linguistic characteristics of Hindi.

IndexTerms - Natural Language Processing, Machine Learning, Artificial Intelligence

INTRODUCTION

Comprising over 350 million speakers[1], Hindi is the most widely spoken language in India and among the most widely spoken languages in the world as well. Consequently, the cultural, political, and social diversity of its speaker base can often result in the spreading of disinformation. The widespread adoption of digital media has further exacerbated the spread and impact of disinformation. A report by the World Economic Forum[2] suggests that at the time of writing, India is the country at the highest risk of disinformation, underlining the necessity of developing systems to ensure the integrity of the information shared.

As such, this paper seeks to develop a comprehensive, end-to-end fact-checking pipeline that is novelly applied to the Hindi language. The system leverages Hindi Wikipedia as a source for evidence retrieval, as it remains a dynamically updating repository of reliable, language-specific content. The pipeline combines innovations in Named Entity Recognition (NER), sentence similarity, and Natural Language Inference (NLI).

While the subjective nature of interpersonal messages posits a fundamental lack of a 'perfect truth' to the validity of statements provided, Wikipedia's usage of peer-validated content has enabled it to develop a comprehensive collection of knowledge, while maintaining the validity of information to a large degree. Therefore, it can be used as a reliable knowledge base for Fact-Checking Systems to reference. Wikipedia has been frequently used in such a scenario in the English-speaking domain, with the FEVER dataset[3] notably propagating growth in the field. From a claim identified, models are incorporated to locate relevant articles and sentences, providing the text necessary for claim verification.

Although Wikipedia has often been cross-referenced in fact-checking models in the English language, limited work has been undertaken in the Hindi domain. Current models comprise sub-tasks such as fact extraction, text-based legitimacy identification, and individual claim verification. [4-7] The lack of end-to-end models presents a void between the intentions and current research of Automated Fact-Checking models. This paper intends to bridge such a gap, incorporating Hindi Wikipedia as an external source for fact verification of text messages.

RELATED WORKS

State-of-the-Art English Fact Verification Methods

Established approaches to the problem of Fact Verification in the English language look to determine the veracity of an input claim, with certain models linking justifications through highlighted sentences. Models generally make use of either graph-based or transformer-based approaches.

Graph-based models, such as those proposed by Zhong et al.[8], use knowledge or linguistic graphs to map the relationship between a claim and its supporting evidence. These graphs are often generated through techniques such as dependency parsing or entity linking. Following the construction of the graph, models such as Graph Convolutional Networks (GCNs) are applied to propagate information and determine its veracity. An attention mechanism is then used to identify subgraphs or nodes with the highest relevance and evaluate the claim's veracity. However, generating and using these graphs requires large corpora of annotated data and sophisticated graph generation methods, which pose challenges due to limited training data and computational complexity.

Transformer-based models make use of NLI architectures, leveraging pre-trained language models such as BERT or RoBERTa[9]. Through the use of transfer learning, the final classification head is fine-tuned to the particular task. The self-attention mechanism implemented in transformers aids in capturing contextual nuances between the claim and the evidence. While transformer-based models generally outperform graph-based ones, they may lack the high-level interpretability offered by structured graph representations.

Hindi NLP

The adoption of multilingual models and transfer learning techniques has been a driving factor for advancements in Hindi NLP tasks. One such example is the application of mBERT (Multilingual BERT), a transformer-based model that supports over 100 languages. By fine-tuning mBERT to emotion recognition problems[10], it has shown remarkable results in accurately predicting emotions from Hindi text, achieving a 91.84% balanced accuracy on test data.

Furthermore, the introduction of datasets such as the HiNER[12] corpus for Named Entity Recognition (NER) has resulted in the growth of training data for models to identify entities in Hindi text. As the application of multilingual models like mBERT continues to provide advantages in Hindi NLP tasks, they can enable more accurate applications in text summarization, emotion analysis, and text generation.

Hindi Detection Methods

The literature in Hindi fact verification exhibits considerable diversity, with various approaches to claim matching. Baisla et al.[5], for instance, employ word embeddings in conjunction with an evaluation of a variety of statistical ML algorithms. In contrast, Sharma et al[6]. make use of transformer models in order to predict the veracity of claims without the use of external verification sources, instead relying solely on subtle differences in language to indicate a claim's truthfulness.

Disentangled Attention For NLI

One of the more recent advancements in transformer-based models is DeBERTa (Decoding-enhanced BERT with Disentangled Attention), which presents a significant improvement over previous models like BERT and RoBERTa. DeBERTa[14] introduces the disentangled attention mechanism, where words in a sequence are represented by two separate vectors encoding their content and position. This disentangled attention framework allows for better handling of syntactic and semantic information, which proves crucial in tasks such as fact verification.

The second innovation is the enhanced mask decoder, which incorporates absolute positional information during pre-training. This method improves masked token prediction, a core task in transformer-based pre-training, helping the model to generalize during fine-tuning on downstream tasks. Empirically, these techniques have significantly enhanced DeBERTa's performance, leading to state-of-the-art results on several benchmarks, including surpassing human performance on SuperGLUE. Given its strength in generalization and its ability to capture intricate linguistic patterns, mDeBERTa, the multilingual variant of DeBERTa, was chosen for this fact-checking pipeline.

DATASETS

FEVER

The Fact Extraction and Verification (FEVER)[3] dataset, developed by Thorne et al., is a pioneering dataset designed explicitly for the Fact Verification task, utilizing English Wikipedia as its knowledge base. The dataset comprises 185,445 claims, each generated through the modification of sentences sourced from Wikipedia articles. These claims are categorized into three classes by annotators: NOT ENOUGH INFO, SUPPORTS, or REFUTES, depending on the availability and nature of evidence supporting or contradicting the claim.

As seen in Table 1, each claim in the FEVER dataset is accompanied by ground truth annotations that include references to the specific Wikipedia sentences used to justify the claim's classification. This explicit linkage between claims and their evidentiary basis makes FEVER a valuable resource for developing and evaluating fact verification models in the English language.

Claim	Evidence	Label
"Telemundo is an English-language television network."	"Telemundo is an American Spanish-language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises."	REFUTES
"James VI and I was a major advocate of a single parliament for Scotland and England."	"He was a major advocate of a single parliament for England and Scotland."	SUPPORTS

Table 1: Examples of claims from the FEVER dataset. Each data point consists of a claim, supporting evidence from a Wikipedia sentence, and a corresponding label.

However, despite its extensive coverage and structured annotations, the FEVER dataset's focus on English-language claims limits its usability in the context of other languages, such as the current model's task centered on Hindi fact verification.

Stancesaurus

Stancesaurus[15] is a multilingual dataset that includes 20,000 tweets across Hindi, Arabic, and English, each containing a distinct claim. These claims are manually annotated by experts, referencing reliable fact-checking sources. The annotations classify each tweet into one of five categories: IRRELEVANT, SUPPORTING, REFUTING, DISCUSSING, or QUERYING, depending on the stance taken in relation to the claim. As seen in Table 2, when combined into the composite dataset, these labels are simplified to YES, NO, or UNCERTAIN.

Claim	Evidence	Label
"Supreme Court orders that marriage of Muslim man and Hindu woman is no longer possible (translated from Hindi)"	"Under the Special Marriage Act, 1954, a Hindu woman can marry a Muslim man, and their marriage will be valid (if they meet the conditions under the Act)."	REFUTING
"Raid at Tirupati temple priest's house, 128 kg gold found (translated from Hindi)"	"Out of 16 priests of Tirupati temple, 128 kg gold, 150 found."	SUPPORTING

Table 2: Examples of claims from the Stancesaurus dataset. Similar to FEVER, each data point includes a claim, evidence, and a final label simplified to YES, NO, or UNCERTAIN.

The Stancesaurus dataset is notable due to its use of several languages. The introduced linguistic and informational diversity is critical in ensuring models making use of the dataset are exposed to a wide array of cultural contexts, making them more relevant to the real-world scenarios they will be exposed to.

CLAIM DETECTION AND MATCHING FOR INDIAN LANGUAGES

Though smaller in scale with 4,096 claims, this dataset[16] offers a unique focus on Indian languages. The claims are derived from diverse sources, such as text messages and social media posts, encompassing a variety of languages, including Bengali, English, Hindi, Malayalam, and Tamil.

Claim	Label
"Is Yogi Adityanath trying to lure prospective voters with money? Fact Check (translated from Bengali)"	NO
"Did BJP IT cell have prior knowledge about ANI's story on Rahul Gandhi and Twitter Bots?"	YES

Table 3: Examples of data points in the claim detection and matching for the Indian languages dataset. Each data point includes a claim and a majority label: YES, NO, or PROBABLY.

As referenced in Table 3, each data point in the dataset includes a fact that has been shared, which is then annotated by three independent evaluators. These annotations are consolidated into one of three categories: YES, NO, or PROBABLY, representing the consensus on the veracity of the claim. Despite its smaller size, this dataset is valuable for developing and testing models tailored to Indian perspectives and topics of interest, addressing the need for culturally and linguistically relevant data in the context of fact verification.

DATASET COMPILATION AND METRICS

The scarcity of Hindi-specific data necessitates the translation and compilation of existing datasets to create a more comprehensive resource for Hindi fact verification.

To achieve this, 20,000 claims from the FEVER dataset were selected to avoid an overrepresentation of English-dominated topics. Moreover, the whole of the Indian language claim detection dataset was added for its highly valuable cultural and linguistic relevance. Table 4 illustrates the allocation of data across the train, test, and cross-validation datasets. The entirety of the Stancesaurus dataset was also included due to its multilingual nature.

To alleviate compatibility concerns, claims were converted into a binary format, with False labeled as 0 and True as 1. While making use of such a binary classification drastically simplifies the integration process, it poses certain challenges. As an example, nuance in stances such as "PROBABLY" or "DISCUSSING" is ignored and consolidated into an output which may not be as valid in real-life contexts.

Non-Hindi claims were translated using the EasyMT library, which ensured consistency and accuracy across the compiled dataset. The final dataset demonstrates a relatively even split between claims annotated as True and False.

Dataset Split	True Claims	False Claims	Total
Training	18,445	18,203	36,648
Validation	2,156	2,134	4,290
Test	2,187	2,155	4,342
Total	22,788	22,492	45,280

Table 4: Complied data across training, testing, and validation sets. The dataset includes both original Hindi claims and their translations. Approximately 85% of the claims are allocated to the training set, with the remaining claims split equally between the validation and test sets.

METHODS

The proposed pipeline makes use of a series of models, each with distinct objectives. The method draws inspiration from the approach of Trokhymovych et al[17]. but involves adaptations for Hindi fact verification.

The pipeline, as seen in Figure 1, begins by identifying relevant Wikipedia articles using a Named Entity Recognition (NER) model. The NER model extracts key entities from the claim, which are then used to search for relevant articles on Wikipedia. This step ensures that the most pertinent sources are identified for fact verification. Once the articles are found, the Python requests module is then used to extract the full text of each article.

Subsequently, a semantic sentence similarity model is used to assign a score to each sentence in the extracted text, determined by. Once the scores are calculated, they are then ranked by relevancy using the Mean Squared Error (MSE) loss.
The top 100 most relevant sentences are selected and propagated into a modified NLI model. This model processes the evidence and the claim to classify the claim as either supported, refuted, or insufficient information. The decision of the NLI model thus forms the final output of the consolidated pipeline.

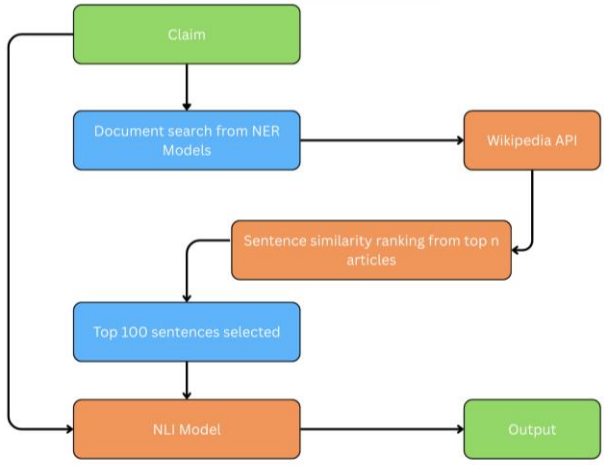

Figure 1: Overview of the model architecture, illustrating the use of NER, sentence similarity ranking, and an NLI model to predict the most probable label for a claim.

NER Identification

As the initial step in the pipeline, a NER-based approach was chosen for its simplicity and effectiveness, providing a baseline level of performance while minimizing the chances of selecting irrelevant articles. The NER model selected was a fine-tuned MuRIL[18] model, which is based on the BERT architecture and pre-trained on a Hindi NLI dataset. Additional training was done using the HiNER dataset, a comprehensive resource for Named Entity Recognition in Hindi.

The BERT-based architecture of MuRIL allows the NER model to achieve a high accuracy in identifying entities within the text. Its deep attention mechanism enables the model to capture context effectively, improving entity detection even in complex or ambiguous sentences.

Procurement using Wikipedia API

Once relevant entities are identified from the claim, each entity is passed to the Wikipedia search engine via its API, retrieving five articles per entity. The extensive data retrieval is undertaken to ensure no critical information is overlooked, for topics may produce multiple related articles.

Ranking Through Semantic Similarity

Once the relevant text has been collected, each sentence is processed by a semantic sentence similarity model, which assigns a relevance score based on its similarity to the original claim. The model employed for this task is a fine-tuned Hindi SBERT model, trained on the Hindi Semantic Textual Similarity (STS)[19] dataset.

The embeddings generated from the retrieved sentences are then compared with the embedding of the original claim using Mean Squared Error (MSE) loss. This approach allows the model to group sentences with similar topics or claims, helping to cluster relevant evidence.

NLI Model

The final step involved a DeBERTa NLI model fine-tuned to the context of the dataset. The model treats the claim as the hypothesis and the retrieved evidence as the premise, then provides an output of either True or False. As mentioned previously, mDeBERTa, a multilingual variant of the DeBERTa model, is pre-trained in 100 languages, making it well-suited for cross-lingual tasks like Hindi fact verification.

RESULTS AND DISCUSSION

As a result of the computational demands of the fact-checking pipeline, arising from the large model sizes involved, experiments were conducted using an Nvidia Tesla P100 GPU. The primary focus during model training was on fine-tuning the Hindi NLI model. As the classification task
To optimize the model's learning process, the Adam optimizer was selected. Adam's adaptive learning rate and faster convergence, especially when handling a high number of model parameters, made it ideal for ensuring time-efficient optimization.

For performance evaluation, multiple metrics were employed to provide a comprehensive assessment of model performance. The F1 score was chosen as the primary metric to ensure a balanced consideration of both true positives and true negatives, crucial for fact-checking where precision and recall are equally important. Precision and recall were also measured individually to understand the model's performance in avoiding false positives and false negatives, respectively. Finally, accuracy was reported to provide an overall measure of correct predictions across all classes.

Results and Analysis

Model	Training F1	Test F1	Training Precision	Test Precision	Training Recall	Test Recall	Training Accuracy	Test Accuracy
mDeBERTa+Pipeline	0.866	0.854	0.871	0.859	0.861	0.849	0.868	0.856
BERT-based classifier	0.758	0.737	0.764	0.743	0.752	0.731	0.761	0.739
WikiCheck (translated)	0.715	0.683	0.721	0.689	0.709	0.677	0.718	0.685
SVM	0.672	0.632	0.678	0.638	0.666	0.626	0.675	0.634

Table 5: Performance metrics for varying models on the compiled training and test datasets. The pipeline-based model outperformed other common alternatives across all metrics.

Results from Table 5 indicate that the pipeline achieved the highest scores across all metrics on both the training and test datasets, demonstrating superior performance in comparison to others seen in existing literature. The mDeBERTa+Pipeline model achieved an F1 score of 0.866 on the training set and 0.854 on the test set. The difference between the training and test further highlights how the model is well-regularized and avoids overfitting.

In comparison, the BERT-based classifier, while performing reasonably well, scored lower across all metrics. The gap between training and test scores indicates a moderately higher level of overfitting. Despite this, the BERT model still demonstrated competitive performance, likely due to its ability to use pre-trained knowledge to handle the fact-checking task.

The WikiCheck model, albeit with inputs and outputs translated from English, achieved the third-highest performance. The larger drop in performance from training to the test set likely arose from challenges associated with language translation and model adaptation. This model's lower performance serves as a reminder of the need for models built specifically for Hindi.

The SVM model performed the weakest across all metrics, with significant gaps between training and test performance. This suggests that traditional machine learning models like SVM struggle with the complexity of the fact-checking task compared to transformer-based models. The substantial difference between training and test scores indicates a higher degree of overfitting.

CONCLUSION

This paper presents a novel Hindi fact-checking pipeline designed to tackle the growing challenge of misinformation in the Hindi-speaking domain. The model innovates by utilizing Wikipedia as a primary source of evidence, along with the use of fine-tuned approaches to maximize accuracy along every step of the pipeline. By building a comprehensive pipeline that integrates Named Entity Recognition (NER), semantic sentence similarity, and a Natural Language Inference (NLI) model, the system provides an end-to-end solution for verifying claims. Experimental results demonstrate superior performance across multiple evaluation metrics, achieving an F1 score of 0.854 on the test set, significantly outperforming traditional machine learning approaches within the language.

LIMITATIONS

While the consolidated pipeline remains largely effective, it nonetheless faces several limitations that could affect its performance when applied to real-world contexts.

Firstly, the pipeline's reliance on Hindi Wikipedia as a primary data source remains a major challenge. Despite its cultural relevancy, Hindi Wikipedia is fundamentally far smaller in scope compared to its English counterpart. As such, when dealing with intricate claims, the lack of evidence may result in inaccuracies.

The use of sentence similarity algorithms has its own limitations. While these models can determine how semantically similar two sentences are, it must be noted that high similarity does not always imply contextual relevance. This can lead to suboptimal evidence selection, where irrelevant information is ranked higher.

Finally, the binary classification approach, while simplifying the task, reduces the nuance of claims in intermediate categories such as "partially true". Once again, this limitation affects the model's ability to handle complex claims that cannot be easily categorized as simply true or false.

ETHICAL CONSIDERATIONS

In a period where the spread of misinformation remains impactful yet burgeoning, the proposed models address one of society's biggest issues. The proposed model's ability to verify the integrity of a claim directly serves as a tool for solving one of society's biggest issues. At the same time, however, the model's positioning as an authoritative "fact checker" placed it with a significant responsibility of maintaining high accuracy. Each error has the risk of further propagating misinformation instead of preventing it. Furthermore, the model's output lacks the nuances of capturing half-truths, or facts which are likely true but cannot be verified. These challenges highlight the imperative need to continuously improve the model to ensure it matches the ethical responsibilities of its task.

ACKNOWLEDGMENT

I am deeply grateful to Dr. Weiwei Sun and Xiaoxi Luo for their invaluable guidance and support throughout my research. I would also like to extend my sincere thanks to the Design and Computer Science departments at my school for fostering my curiosity and passion for the field of Machine Learning.

REFERENCES

1. Office of the Registrar General & Census Commissioner, India. *C-17: Population by Bilingualism and Trilingualism*; 2011.
2. World Economic Forum. *Global Risks Report 2024*. World Economic Forum. https://www.weforum.org/publications/global-risks-report-2024/.
3. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. *FEVER: a Large-scale Dataset for Fact Extraction and Verification*. ACLWeb. https://doi.org/10.18653/v1/N18-1074
4. Praseed, A.; Rodrigues, J.; Thilagam, P. S. Hindi Fake News Detection Using Transformer Ensembles. *Engineering Applications of Artificial Intelligence* 2023, *119*, 105731. https://doi.org/10.1016/j.engappai.2022.105731.
5. Sohail Baisla; Aggarwal, M.; Bansal, P.; Malik, K. Hindi Fake News Fact Checker Using Machine Learning and Deep Learning. 2023, *686*, 421–433. https://doi.org/10.1007/978-981-99-2100-3_33.
6. Sharma, R.; Arya, A. LFWE: Linguistic Feature Based Word Embedding for Hindi Fake News Detection. *ACM Transactions on Asian and Low-Resource Language Information Processing* 2023, *22* (6), 1–24. https://doi.org/10.1145/3589764.
7. Tejashree Ghude; Chauhan, R.; Krushna Dahake; Atharv Bhosale; Tushar Ghorpade. N-Gram Models for Text Generation in Hindi Language. *ITM Web of Conferences* 2022, *44*, 03062–03062. https://doi.org/10.1051/itmconf/20224403062.

8. Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; Yin, J. Reasoning over Semantic-Level Graph for Fact-Checking. *Association for Computational Linguistics* 2020, 6170–6180. https://doi.org/10.18653/v1/2020.acl-main.549.
9. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* 2019.
10. Kumar, T.; Mehul Maharishi; Sharma, G. Emotion Recognition in Hindi Text Using Multilingual BERT Transformer. *Multimedia Tools and Applications* 2023, *82* (27), 42373–42394. https://doi.org/10.1007/s11042-023-15150-1.
11. Vijay, S.; Rai, V.; Gupta, S.; Anshuman Vijayvargia; Sharma, D. M. Extractive Text Summarisation in Hindi. *IEEE* 2017, 318–321. https://doi.org/10.1109/ialp.2017.8300607.
12. Murthy, R.; Bhattacharjee, P.; Sharnagat, R.; Khatri, J.; Kanojia, D.; Bhattacharyya, P. *HiNER: A Large Hindi Named Entity Recognition Dataset*. arXiv.org. http://arxiv.org/abs/2204.13743 (accessed 2025-03-31).
13. Tushar Abhishek; Shivprasad Sagare; Singh, B.; Sharma, A.; Gupta, M.; Varma, V. XAlign: Cross-Lingual Fact-To-Text Alignment and Generation for Low-Resource Languages. *Companion Proceedings of the The Web Conference 2018* 2022, 171–175. https://doi.org/10.1145/3487553.3524265.
14. He, P.; Liu, X.; Gao, J.; Chen, W. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv.org. http://arxiv.org/abs/2006.03654 (accessed 2025-03-31).
15. Zheng, J.; Baheti, A.; Naous, T.; Xu, W.; Ritter, A. *Stanceosaurus: Classifying Stance Towards Multilingual Misinformation*. arXiv.org. http://arxiv.org/abs/2210.15954 (accessed 2025-03-31).
16. Kazemi, A.; Garimella, K.; Gaffney, D.; Hale, S. A. *Claim Matching Beyond English to Scale Global Fact-Checking*. arXiv.org. https://arxiv.org/abs/2106.00853 (accessed 2025-03-31).
17. Trokhymovych, M.; Saez-Trumper, D. *WikiCheck: An end-to-end open source Automatic Fact-Checking API based on Wikipedia*. arXiv.org. http://arxiv.org/abs/2109.00835 (accessed 2025-03-31).
18. Khanuja, S.; Bansal, D.; Mehtani, S.; Khosla, S.; Dey, A.; Gopalan, B.; Kumar, M. D.; Aggarwal, P.; Teja, N. R.; Dave, S.; Gupta, S.; Bose, C.; Subramanian, V.; Talukdar, P. *MuRIL: Multilingual Representations for Indian Languages*. arXiv.org. http://arxiv.org/abs/2103.10730 (accessed 2025-03-31).
19. Agarwal, D.; Mujadia, V.; Mamidi, R.; Sharma, D. Semantic Textual Similariity for Hindi. 2017.