CSE508: Information Retrieval

LSD for Indian Languages :Lexical Text Simplification and Detoxification

Aanya Trehan	Abhimanyu Bhatnagar	Aniket Goel	Apoorva Arya	Naman Kaushik	Sejal Kardam
2020419	2020273	2020281	2020032	2020088	2020467

1. Problem Statement

Our group aims to build a system to replace complex and difficult to understand words in a text with less complex words. However, developing effective lexical simplification techniques remains a challenging task due to the complexity of language and the nuances of meaning that can be lost in the simplification process. Therefore, we aim to develop a novel technique that can accurately and effectively simplify the text while maintaining the meaning and preserving the intended tone and style of the original text. We intend to make it more inclusive by incorporating different languages like Hindi and if possible a few regional languages. Additionally, we would also be working towards detecting offensive text to detoxify it so that it can be viewed by a large audience without repercussions. This would require developing sophisticated algorithms and models that can accurately identify offensive text-based for which we would use methods like sentiment analysis.

2. Importance of the Problem

Many individuals with different cognitive and linguistic abilities struggle to understand the complexity of texts, which limits their access to information and impedes their ability to participate in daily activities. Lexical Text Simplification aims to address this challenge by automatically simplifying the language and structure of text while preserving the original meaning and intent of the message. Moreover, in today's society, there is a growing concern over the increase in hate speech which introduces the need for the detoxification of the text.

3. Related Works

3.1 TextRazor

TextRazor is a natural language processing (NLP) platform that utilizes machine learning algorithms to extract information from unstructured text, including tweets, web pages, and other text-based databases. However, it has limitations, such as being trained only on general text, which means it may not be able to analyze data in specialized domains, like scientific or technical texts. Additionally, it struggles to handle texts with complex sentence structures and has limited language support. Moreover, TextRazor is a paid service, which may not be feasible for users with limited budgets.

3.2 HateSonar

HateSonar is an NLP tool that is specifically designed to detect hate speech in text. It analyses text using ML techniques to find linguistic patterns frequently found in hate speech. It gives false positive results as the models are not appropriately trained [1]. It performs poorly on our random sample with a precision of 0.5 and recall of 0.31 [2]. It provides limited language support and limited customization and is unable to identify some forms of hate speech, such as bullying or threats.

3.3 Perspective API Google

Perspective is a technology developed by Google that utilizes machine learning algorithms to detect abusive comments by scoring the impact of a given text in a conversation. Developers and publishers can use this score to provide feedback to commenters, make it easier for moderators to review comments, or allow readers to filter out offensive language [3]. However, An adversary can subtly modify a highly toxic phrase so that the system assigns a significantly lower toxicity score to it [4]. It faces difficulty in handling sarcasm and irony and difficulty keeping up with rapidly evolving languages and slang used in online conversations.

3.4 Rewordify

Rewordify is a web tool that makes difficult English easier to understand. It replaces hard English words and phrases with simplified forms [5]. This tool is designed primarily to simplify low-frequency difficult words; hence when sentences are made using simple high-frequency words, Rewordify cannot simplify the sentence further [6]. It also struggles in paraphrasing while retaining the context in certain scenarios like academic work with specific scientific language [7].

3.5 Simplish

Simplish is a tool that simplifies by rewriting text using fixed 1000 words of basic English vocabulary. It also tries to explain the difficult words using these basic English words. What makes it unique is the fact that instead of replacing only difficult words or phrases in a text, it replaces all the words [8]. This can lead to a few

problems, such as over-simplification of the text and changing context. Another issue is when the algorithm is unable to replace a term, it substitutes it with its dictionary definition instead. This results in an extremely long output that might be more difficult to understand. Its paid model offers more features, but that is not accessible to everyone.

3.6 Detoxification

The available detoxifying techniques are exclusive to one particular language and are not easily adaptable to other languages. Research shows that multilingual models are capable of transferring styles between different languages. However, evaluations show that they were unable to perform detoxification across many languages [9]. Detoxification methods have been proven to be detrimental to equity since they reduce the usefulness of LMs for the language used by excluded groups [10]. Particularly when it comes to the language used by marginalized groups, detoxification renders language models more susceptible to distribution change [10].

3.7 Lexical Text Simplification

Similarly, there is limited implementation of Lexical Text Simplification in other languages like Indian regional languages. For example, the first Automatic Text Simplification tool for Urdu was developed in 2020 which can only simplify one word [11]. Recent research has also highlighted a few challenges faced like limited high-quality resources and corpus. There is also a need for improved standardized measures of usability, evaluation, and quality of the output [12].

4. Novelty

Our proposed model stands out from existing models due to several unique features. Firstly, it employs a more context-oriented simplification technique that retains the meaning of the sentence, unlike other models that often change it. Secondly, it aims to be more inclusive by incorporating support for Indian regional languages, which are often overlooked by current models. Furthermore, it also includes a grammar correction component, which is not offered by many functional tools. Finally, to enhance the accuracy of the simplification process, we plan to develop a tailor-made metric that outperforms existing automatic metrics and produces simplified text that is not only more user-friendly but also closer to the ground truth. These features combined make our model novel and different, and potentially a valuable tool for simplifying complex text for a broader audience.

5. Proposed Solution *Techniques and Evaluation*

The proposed project aims to address a critical issue in natural language processing (NLP) through the development of an innovative mechanism. Specifically, the project aims to create a framework that can undertake two fundamental NLP tasks: text simplification and detoxification.

To accomplish the task of text simplification, the proposed mechanism will utilize publicly available datasets that contain sentence pairs of complex and simplified English text, including WikiLarge, Newsela, ASSET, and ParaCrawl. The project will implement a two-pronged approach consisting of a language model for text generation and an evaluation metric for readability assessment. The state-of-the-art BERT model and the LED model will be utilized for language modeling. In addition, a range of established automatic readability metrics, such as the Flesch-Kincaid score, Gunning-Fog Index, and Coleman-Liau index, will be applied. To further enhance the efficacy of the model, custom-tailored metrics that better capture the simplification requirements will also be developed. Human evaluation will also be incorporated to validate the accuracy of the system.

Furthermore, since the proposed mechanism aims to simplify text for the general public, the team will take additional measures to ensure grammatical accuracy. In this regard, the project will integrate a module that automatically detects and corrects any grammatical errors that may occur during the text simplification process.

The proposed mechanism will also undertake the task of text detoxification, which involves identifying and modifying toxic content. The project will employ the VADER sentiment analysis tool to ascertain the presence of toxic language in the input text. If toxic language is identified, techniques such as synonym replacement and word deletion will be employed to remove offending words while preserving the meaning of the text.

This project represents a cutting-edge solution to a critical problem in NLP and is expected to yield significant benefits for a range of applications, including enhancing text accessibility and reducing the spread of harmful language on online platforms.

6. Potential Contributions

All of us would be involved with the data collection, cleaning, model development, and subsequent analysis of the result. Additionally, the potential domain distribution is as follows:

AANYA	Dev, ML	APOORVA	*ML, NLP*
ABHIMANYU	*ML, NLP*	NAMAN	Dev, ML
ANIKET	*ML, NLP*	SEJAL	Dev, ML

7. References

[1] Hironsan, “hatesonar: Hate Speech Detection Library for Python.”. Available: https://github.com/Hironsan/HateSonar

[2] S. Zannettou, M. Elsherief, E. Belding, S. Nilizadeh, and G. Stringhini, “Measuring and Characterizing Hate Speech on News Websites,” in 12th ACM Conference on Web Science, Southampton United Kingdom, Jul. 2020, pp. 125–134. doi: 10.1145/3394231.3397902.

[3] “Perspective API - How it works.” https://perspectiveapi.com/how-it-works/ .

[4] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, “Deceiving Google’s Perspective API Built for Detecting Toxic Comments.” arXiv, Feb. 26, 2017. doi: 10.48550/arXiv.1702.08138.

[5] “Rewordify.com | Understand what you read.” https://rewordify.com/index.php .

[6] “Rewordify.com | Help Center.” https://rewordify.com/helpwhatcannotdo.php .

[7] P. Athanasiadou, G. Andreou, and E. Gana, “ICT and specific learning disabilities: A proposition for the use of the software Rewordify in the foreign language learning by students with reading comprehension difficulties,” Διεθνές Συνέδριο Για Την Ανοικτή Εξ Αποστάσεως Εκπαίδευση, vol. 10, no. 3A, Art. no. 3A, 2019, doi: 10.12681/icodl.2298.

[8] “Need to simplify/summarize text online?” https://www.simplish.org/\.

[9] D. Moskovskiy, D. Dementieva, and A. Panchenko, “Exploring Cross-lingual Text Detoxification with Large Multilingual Language Models.,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Dublin, Ireland, May 2022, pp. 346–354. doi: 10.18653/v1/2022.acl-srw.26.

[10] A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein, “Detoxifying Language Models Risks Marginalizing Minority Voices.” arXiv, Apr. 13, 2021. doi: 10.48550/arXiv.2104.06390.

[11] N. H. Qasmi, H. B. Zia, A. Athar, and A. A. Raza, “SimplifyUR: Unsupervised Lexical Text Simplification for Urdu,” in Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 3484–3489. [Online]. Available: https://aclanthology.org/2020.lrec-1.428

[12] S. Stajner, “Automatic Text Simplification for Social Good: Progress and Challenges,” in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, Aug. 2021, pp. 2637–2652. doi: 10.18653/v1/2021.findings-acl.233.