Response to the Reviews

Title: A Rule-based Recursive Lemmatisation Algorithm with POS tagging for Plagiarism Detection in Nepali texts

Manuscript Reference Number:
IEEE TCST-2019-0001

Authors:
Ayush Kumar Shah

Date: October 27, 2020

Message from the Authors

Dear Reviewers,

We thank you for your constructive comments, which have allowed us to improve the quality of the manuscript. We have addressed the comments and incorporated your valuable suggestions in the revised manuscript, in particular highlighting the key contributions and significance of this work.

We have made the following major changes to the original manuscript:

1. **Change in title:** We have changed the title by removing the adjective novel and adding terms to better represent the contributions in the paper.
2. **Addition of contributions in Abstract**: We have added four major contributions in the abstract. We have also added the evaluation result scores to quantify the significance of the proposed method well ahead.
3. **Improvement in the Introductions section:** The introduction section has been modified to get into the specific challenges involved in pre-processing of Devanagari scripts quickly, removing the unnecessary background. Likewise, several references of works done in stemming of Devanagari scripts have been added.
4. **Addition of Discussions and Comparisons sub-section:** A separate sub-section has been added under the section "VI. Results and Discussions" to discuss about the limitations in existing stemming approaches in Devanagari scripts, the changes adopted in the proposed method from previous rule-based methods, and significance of the approach in detail with comparisons of scores with other methods.
5. **Introduction of new publicly available dataset:** A new publicly available dataset called NEP-PLAG2019v1 has been introduced, which is supposed to contain 10,000 pairs of annotated Nepali news articles with plagiarism class labels. However, this dataset is not available yet and is a work for future.
6. **Addition of Experiments:** The original experiments have been conducted on the newly introduced dataset, consisting of 10,000 pairs of articles instead of small 100 pairs of articles. Further experiments have been added to compare the results of the proposed method with previous similar approaches.
7. **Improvement in visibility of figures:** Figures have been made more clear and large figure has been rendered in a complete separate page for better visibility.

We address each comment separately in the following detailed response. The comments we received are boxed, and our responses are written following each comment. All page and reference numbers in our response are based on the revised manuscript, unless otherwise stated. The page and reference numbers mentioned in the reviewers' comments are kept intact and are based on the original manuscript. We look forward to hearing from you and hope that you find the revised manuscript satisfactory.

Sincerely,
Ayush Kumar Shah

Response To Reviewer #1 (Akib Shahriyar)

Reviewer Comment

> The authors have developed a stemming algorithm and tested it on a relatively small dataset (100 pairs of Nepali news articles).

Response

We have introduced a new dataset called NEP-PLAG2019v1, consisting of 10,000 pairs of articles. Please refer to the 'Introduction to new publicly available dataset' point in the major changes summary.

Reviewer Comment

> The authors have claimed that no prior work has been done on the pre-processing of Devanagari scripts without any substantial literary reference. The claimed contributions are not completely novel. Previous work has already been done on developing a stemming algorithm for Devanagari Scripts. The experimental results are presented without comparing them with any similar method/prior work.

Response

We have realized the mistake pointed by the reviewer in regards to our claim that no prior work has been done. We have added references to prior works on the preprocessing of Devanagari scripts in the Introduction section. Please refer to the 'Improvement in the Introductions section' point in the major changes summary.

New experiment results have been added to compare them with prior works in the C. Discussions and Comparisons sub-section under VI. Results and Discussions section.

Reviewer Comment

> The explanation of the methods is insufficient for a reader to replicate it thoroughly. Specifically, the dataset/corpora are not publicly available. Moreover, authors have manually annotated the dataset and also collected stop-words from various online locations. There are no references in the paper that could lead any reader to those resources. These points should be taken care of by the authors.

Response

Further explanations of the methods have been added by creating a more sub-sections. The Methodology section in the original manuscript has been replaced by steps of the methods so that they

can be explained in detail.

The plagiarism dataset, the code and all the used corpora including stop-words have been provided publicly with links included in the footnotes so that the methods can be replicated by the readers easily.

Reviewer Comment

> The Methodology Subsection and Stemming and Lemmatization algorithm subsection need a complete overhaul. The text size in the figures is too small for the usual reading. They should be updated. Also, there are some typographical errors and it needs proofreading.

Response

Please refer to the previous response regarding the sub-division of methodology section. The figures have been updated to increase visibility and the large figure containing the entire algorithm flow diagram has been placed in a separate one column page.

All the typographical errors have been checked and corrected.

Response To Reviewer #2 (Murtaza Tamjeed)

Reviewer Comment

> Preprocessing a foreign language text is a great contribution; however, for plagiarism detection, there are many more factors to consider. Presence of the common base forms of the words in two articles may not guarantee plagiarism. Some other factors that need to be considered for plagiarism detection might include, context, sentence level similarity, and document level similarity.

Response

We agree that there are several factors to consider for plagiarism detection like the ones mentioned by the reviewer. We have included them as part of the limitations and future work in the VII. Conclusion section. The current approach has produced fair results for the dataset being used, hence we have retained the main approach in the paper.

Reviewer Comment

> Also, the statistical measures reported are from 20 pairs of manually annotated articles. These measures might not hold true on real plagiarized texts. This work would have been improved if the author had trained and tested the algorithms on larger set of real plagiarized and non-plagiarized text.

Response

We have introduced a new dataset called NEP-PLAG2019v1, consisting of 10,000 pairs of articles and experiments have been added on the new dataset. Please refer to the 'Introduction to new publicly available dataset' and 'Addition of Experiments' points in the major changes summary.

Response To Reviewer #3 (Dingrong Wang)

Reviewer Comment

> The work is very solid and impressive. But I think in order for this paper to be published, you still need some revision. For example, you should explain more detailly about how to do lemmatization, such as why there still exists a check for suffix and prefix after previous two check about prefix and suffix?

Response

Further explanations of the methods have been added by creating a more sub-sections. The Methodology section in the original manuscript has been replaced by steps of the methods so that they can be explained in detail.

The specific questions regarding the check for suffix and prefix after previous two checks have been addressed in the algorithm description in point 6 of the sub-section E. Stemming and Lemmatization Algorithm. Likewise, the code and datasets have been made publicly available so that the methods can be replicated by the readers easily.

Reviewer Comment

> And another question, why do you choose to recombine suffix, but not prefix, is there a combination problem here? Besides, the Jaccard similarity you talk about in the end just concerns the vocabulary info, not the frequency info, right? I think you should explain more about these questions for the reader to get a better understanding.

Response

Please refer to the previous response regarding the addressing of questions in the algorithm description. This question has also been addressed now.

The issues with Jaccard similarity and the reason for discarding it for classification of texts has been added at the end of the B. Experiment sub-section of VI. Results and Discussions section.

Reviewer Comment

> And if you want to improve your result, I think you should try some deep learning method in NLP to dig out more semantic and contextual information to construct feature vector and further compute the similarity extent.

Response

We agree that there are several deep learning methods which can provide better results for plagiarism detection. We have added these methods as part of the limitations and future work in the VII. Conclusion section. The current approach has produced fair results for the dataset being used, hence we have retained the main approach in the paper. However, we intend to use deep learning approaches in future for increasing the performance accuracy in more complex dataset.