Review for Journals

Date: 11th October 2020

Manuscript ID:

Title: A Novel Rule-based Recursive Stemming Algorithm for Plagiarism Detection in Devanagari Scripts.

Author(s): Ayush Kumar Shah

• **Goals and Contributions:**
i. Do the authors clearly state the research goals of the work?

The authors have stated the research goals of the work and also described its importance with relevant scenarios.

ii. Does the paper clearly indicate what the contributions are?

The last paragraph of the Introduction section states the contribution of the authors.

iii. Are the claimed contributions original and significant in terms of
o Novel methodology?

The claimed contributions are not completely novel. Previous work has already been done on developing a stemming algorithm for Devanagari Scripts.

o New applications?

The application of the stemming algorithm for Plagiarism detection in Devanagari Scripts is not completely new. Prior work has been done to detect plagiarism in Devanagari Scripts (Nepali Language) by applying Neural Networks.

iv. Does the paper describe the methods in sufficient detail for readers to replicate the work?

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.

• **Evaluation:**
i. Do the authors carefully evaluate the approach?

The authors have utilized two evaluation metrics, Cosine similarity, and Jaccard similarity. Both of them are lacking proper reference to the root article. Moreover, they have not compared the proposed method with any earlier work.

ii. Does the paper include systematic experiments, a careful theoretical analysis, or give evidence of generality?

The authors tried to follow the systematic experiments but lacked in some areas. Critical analysis of their work and the significance of their work is not clearly represented in the paper.

- **Discussion:**

i. Does the paper discuss relevant earlier works, noting similarities, differences and progress?

The authors have claimed that no prior work has been done on the pre-processing of Devanagari scripts which lacks substantial literary proof. No related work has been discussed in the Introduction section either.

ii. Does it discuss the limitation of the approach as well as its advantages?

They discuss some limitations of their approach but overlooks some major areas like dataset preprocessing steps, lack of access to the corpus, lack of comparison with related works.

iii. Does it consider the implication of the work and outline direction for future work?

Yes, the authors have discussed their work's future implications and also shared the direction they would follow next.

- **Presentation:**

i. Is the paper properly organized and well written?

Although the writing quality of the authors is good, the paper lacks proper organization. Specifically, 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.

ii. Is the paper grammatically correct and free of spelling
errors?

There are some typographical errors and it needs proofreading.

iii. Does it use standard terminology?

Yes, the paper use standard terminology.

- **Detailed Comments:**

This article presents a rule-based stemming algorithm for Devanagari Scripts to detect plagiarism in Nepali documents. The authors have developed a stemming algorithm and tested it on a relatively small dataset (100 pairs of Nepali news articles). The experimental results are presented without comparing them with any similar method/prior work. The authors have claimed that no prior work has been done on the pre-processing of Devanagari scripts without any substantial literary reference. The presentation of the work is somewhat complicated and it is difficult to follow. Therefore, the article needs a complete revision to consider it for publication.

- **Recommendation:**

i. The paper could be published in its current form.

ii. The paper could be published after minor revision:

- o Another round of review is needed.
- o No review is needed.

iii. [✓ The paper requires major revision for further consideration.]

iv. The paper is not suitable for publication in this journal.

Review for Conferences

Date: 11th October 2020

Manuscript ID:

Title: A Novel Rule-based Recursive Stemming Algorithm for Plagiarism Detection in Devanagari Scripts.

Author(s): Ayush Kumar Shah

Reviewer's Recommendations:
i. Writing (choose one)

- **o** Not readable
- **o** [✓ Major improvement needed]
- **o** Minor improvement suggested
- **o** Well written

ii. Novelty (choose one)

- **o** Original
- **o** Somewhat interesting
- **o** [✓ Borderline]
- **o** Been there, done that

iii. Suitability (choose one)

- **o** Very related
- **o** [✓ Limited interests]
- **o** Not suited

iv. Reviewer's Expertise (choose one)

- **o** Expert
- **o** Knowledgeable
- **o** [✓ Passing interests]
- **o** Not my cup of tea

v. Recommendation (choose one)

- **o** Absolute reject
- **o** [✓ Reject if there is no space]
- **o** Accept if there is space
- **o** Absolute accept

• Reviewer's detailed comments: *strength, weakness, suggestion*

i. **Comments for the Authors:**

This article presents a rule-based stemming algorithm for Devanagari Scripts to detect plagiarism in Nepali documents. The authors have developed a stemming algorithm and tested it on a relatively small dataset (100 pairs of Nepali news articles).

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.

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.

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.

ii. **Comments for the Program Committee** (will be kept confidential and NOT released to the authors)

Although the writing quality of the authors is good, the paper lacks proper organization and also needs major changes to multiple sections. The work presented here is interesting but is not completely novel. The presentation of the work is somewhat complicated and it is difficult to follow. Therefore, the article needs a major revision to consider it for publication and can be rejected if there are no spaces.