Project Title: Misinformation Detection for South African low resource languages

Application name: Seani Rananga
Affiliation: Lecturer of Computer Science at the University of Pretoria in South Africa
Email: seani.rananga@up.ac.za

Keywords: Misinformation detection, low resource languages, South Africa

1. Description of a resource related to the topics of the training school:

Misinformation detection in South African languages can benefit from the Universal Dependencies (UD) Treebank in several ways:

Training Data:
The UD Treebank provides annotated data for various languages, including some South African languages like Zulu, Xhosa, and Afrikaans. This annotated data can be used to train machine learning models for misinformation detection. By leveraging the syntactic structures and grammatical relations annotated in the UD Treebank, models can better understand the natural language patterns and linguistic nuances specific to South African languages, thus improving the accuracy of misinformation detection.

Feature Extraction:
The syntactic information encoded in the UD Treebank can serve as valuable features for detecting misinformation. Features such as dependency relations, part-of-speech tags, and syntactic dependencies can provide important clues for identifying deceptive or misleading content. For example, certain syntactic structures or grammatical constructions may be indicative of biased or sensationalized language commonly found in misinformation.

Cross-Lingual Transfer Learning:
Since the UD Treebank provides annotations for a wide range of languages, including some non-South African languages, it can facilitate cross-lingual transfer learning. Models trained on annotated data from one language can be fine-tuned or adapted to perform misinformation detection in South African languages with limited labeled data. This is particularly useful for languages with scarce resources, as it allows for knowledge transfer from resource-rich languages to resource-poor languages.

Semantic Understanding:
Misinformation detection often requires a deep understanding of the semantics and context of the text. The dependency annotations in the UD Treebank capture the semantic relationships between words in a sentence, which can help models to better grasp the intended meaning and detect subtle forms of misinformation. For instance, identifying contradictions or inconsistencies in the semantic structure of a sentence can be indicative of misinformation.

Evaluation and Benchmarking:
The UD Treebank provides a standardized framework for evaluating and benchmarking NLP models. Researchers and developers working on misinformation detection for South African languages can use the UD Treebank as a benchmark dataset to evaluate the performance of

their models and compare them against state-of-the-art approaches. This facilitates reproducibility and ensures consistency in the evaluation process.
Overall, leveraging the Universal Dependencies Treebank can enhance the effectiveness and robustness of misinformation detection systems for South African languages by providing valuable resources, features, and evaluation benchmarks tailored to the linguistic characteristics of these languages.

2. Explanation how the participation in the training school will be useful for the project:

Participating in a training school can offer several benefits for your project on misinformation detection for South African low-resource languages:

Skill Enhancement:
Training schools often provide intensive workshops and lectures on various topics related to natural language processing (NLP) and machine learning. By participating, you can acquire new skills and deepen your understanding of relevant techniques and methodologies for misinformation detection. This can include topics such as text classification, natural language understanding, feature engineering, and model evaluation, all of which are crucial for building effective misinformation detection systems.

Access to Resources:
Training schools may provide access to resources such as datasets, tools, software libraries, and computational resources that are essential for conducting research in NLP and machine learning. These resources can be valuable for your project, especially in the context of low-resource languages where access to annotated data and specialized tools may be limited.

Networking Opportunities:
Training schools bring together researchers, practitioners, and experts from academia and industry, providing valuable networking opportunities. By interacting with peers and experts in the field, you can exchange ideas, collaborate on research projects, and gain insights into best practices and emerging trends in misinformation detection. Networking can also lead to potential collaborations or partnerships that can enhance the impact and reach of your project.

Feedback and Mentorship:
Training schools often feature mentorship sessions, where participants receive feedback and guidance from experienced researchers and practitioners. This feedback can be invaluable for refining your research methodology, addressing challenges, and improving the quality of your work. Mentorship can also provide you with valuable insights and perspectives from experts who have experience working on similar projects or in related domains.

Research Exposure:
Training schools frequently include poster sessions, paper presentations, and panel discussions where participants can showcase their research and engage with the broader NLP community. Presenting your project at a training school can increase its visibility, attract potential collaborators or stakeholders, and stimulate discussions that may lead to new ideas or research directions. It also provides an opportunity to receive feedback from peers and experts in the field, which can help you refine and strengthen your research contributions.

Overall, participating in a training school can significantly enhance your project on misinformation detection for South African low-resource languages by providing you with the skills, resources, networking opportunities, mentorship, and research exposure necessary to advance your research goals and make meaningful contributions to the field.

3. Open questions related to the project which could be addressed during the brainstorming hackathon
   3.1 What strategies can be employed to ensure the quality and reliability of annotated data, especially when dealing with complex linguistic structures and dialectal variations?
   3.2 What methods are effective for adapting models trained on resource-rich languages to perform well on low-resource languages without extensive labeled data?
   3.4 How can we design feature representations that capture linguistic nuances and cultural context specific to each language?

4. Short statement of the project phase (planning, started, in the process of creation)

   This project has an approved proposal for implementation, and the current focus is on data collection and analysis. As part of this stage, we are collecting data from social media using paid APIs. However, there's a need to attend this summer school to assist with annotation and improve the quality of the collected data.

   We aim to translate this data into South African languages and ensure proper annotation. This step is crucial for effectively detecting misinformation in low-resource languages. By attending the summer school, we hope to gain insights into best practices for annotation and techniques to enhance the quality of our data, ultimately improving the accuracy of our misinformation detection efforts.