Report on the Outcomes of a Short-Term Scientific Mission[1]

Action number: 21129

Grantee name: Edlira Gugu

Details of the STSM

Title: *Practice on Computational Analysis of Opinion - Albanian Case*

Start and end date: 25/02/2024 to 01/03/2024

Description of the work carried out during the STSM

The research *Practice on Computational Analysis of Opinion - Albanian Case* was conducted at the University of Salamanca, Department of Sociology and Communication, at the *Observatorio de Contenidos Audiovisuales* (OCA) from February 26th to February 29th, 2024. Professor Carlos Acila Calderon supervised the practice.

First day 26.02.2024 - *Carlos Acila Calderon*

Presentation with the working group, the members of OCA, aim of the visit, and schedule. Introduction to R and Python, benefits, versatility, extensive libraries, and robust community support.

- Installing interpreters for both R and Python.
- Detailed instructions were provided for installing R and R Studio, as well as Python and Jupyter Notebook, along with any necessary third-party packages.

Introduction of programming concepts and the essential concept for data analysis. The basic programming concepts, which include:

- variables, basic control structures, data structures, object-oriented programming, troubleshooting and debugging;
- how to use Python libraries and especially them to structure medium datasets.

There were practice exercises on how to write code effectively, with a focus on clarity and efficiency by coding exercises and practical demonstrations.

There were also discussions on drafting the survey for the paper: Computational Literacy of Opinion Researchers across Europe – as Academic Output for WG4.

Visit to the Historic Main Building of the University, the Hotel of the University, etc.

Second day 27.02.2024: *Carlos Acila Calderon and Patricia Sánchez Holgado*

Exploratory data analysis (EDA) techniques, understanding their underlying patterns and structures.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Practice in how to perform various statistical analyses, visualize data using graphs and charts and identify trends and outliers.
Statistical modelling techniques, focusing on supervised machine-learning algorithms, like Naive Bayes and Support Vector Machine (SVM).
There was developed a good understanding of how these algorithms can be used to build predictive models by learning from labelled data; training and evaluating supervised machine learning models using Python libraries such as scikit-learn.
There were explored methods for handling text data stored in different formats, such as CSV files or .txt files organized in folders using Python; processing text data using existing packages in Python; techniques for cleaning and pre-processing textual data, including tokenization, lemmatization, part-of-speech, conversion to lowercase, removing special characters or other unnecessary data, etc.
The concept of treating text as data. There was practice:

- how to analyse textual data stored in a data frame format;
- how to feature extraction, the Bag-of-Word (BoW) model to represent word repetition in a document, utilizing a vocabulary of unique words and a frequency measure;
- the TF-IDF technique to generate the word's frequency in a document adjusted for document length and the inverse frequency of the word across a set of documents;
- creating a vector to represent text content, and TF-IDF associates each word with a number expressing its importance;
- utilizing scikit-learn library to implement these steps.

Through practical exercises, there was hands-on experience in how to prepare data to be used to train models based on machine learning algorithms, like Naive Bayes and SVM, for sentiment analysis and opinion mining. There were acquired practical skills in exploratory data analysis, statistical modeling, and text processing using Python.
Visit to the library of the University of Salamanca

Third day 28.02.2024 *William González, Félix Ortega*
Practice in techniques for the automatic analysis of text. Practice in how to leverage NLP libraries and tools in Python, such as NLTK (Natural Language Toolkit) to perform tasks such as tokenization, part-of-speech tagging, parsing, classification, stemming and sentiment analysis etc.
Discussion on models based on research of William for hate speech in islamophobia.
Practice in web scraping for collecting online data; how to use Python libraries like BeautifulSoup and Scrapy to extract data from websites, including structured data from HTML pages and unstructured data from text documents; practice in exploring techniques for scraping network data from social media platforms and other online sources.
Practical sessions in creating web-scraping scripts to extract specific information from websites, setting up automated data collection pipelines, and processing the scraped data for analysis. Practice in how to automate the analysis of textual data and scrape online data from various sources, enabling me to gather relevant data for my research.
Visit to the Radio – Television of the University of Salamanca

Forth day 29.02.2024 *Carlos Acila Calderon*
Practice in-text analysis of Albanian language data, with a human-annotated dataset tailored for social media sentiment analysis in the Albanian language. Practice how to pre-process the text data, including tasks such as tokenization, stemming, and removing stop words, to prepare it for analysis. Preparation of the data to be used to create sentiment analysis and opinion mining models by training machine-learning algorithms; dividing the dataset in training, development, and testing; training in the model using Naive Bayes and SVM algorithm using the training and development dataset; evaluate the model accuracy by using unseeing before data by the model, the test dataset. Practice in performing sentiment analysis and opinion mining for text in the Albanian language; practice in exploring techniques for evaluating the performance of sentiment analysis and opinion mining models and fine-tuning their parameters to improve accuracy and robustness.

Description of the STSM main achievements and planned follow-up activities

The Short-Term Scientific Mission (STSM) *Practice on the Computational Analysis of Opinion - Albanian Case* successfully achieved its planned goals and expected outcomes, contributing significantly to the Action's objectives and deliverables.

The STSM aimed to enhance the researcher's skills in computational analysis of opinion, particularly focusing on text analysis and sentiment analysis. The researcher gained hands-on experience and practical skills in using Python programming language and various libraries for data analysis, including R and NLTK. The STSM successfully covered a wide range of topics, including exploratory data analysis, statistical modelling, text processing, automatic text analysis, web scraping, and sentiment analysis.

The STSM directly contributed to the Action's objective of advancing research in computational analysis of opinion, in transmitting experience, and helping a researcher enhance skills, particularly in the context of analysis of the Albanian language.

By acquiring new methodologies and skills, the researcher has expanded the knowledge base of the Action, providing insights into text analysis techniques and sentiment analysis models.

The practical sessions and discussions during the STSM helped refine existing methodologies and explore innovative approaches to opinion analysis, aligning with the Action's goals.

As the results of the STSM outcomes, the researcher will develop three days' workshop *mentoring schemes* with a special focus on capacity building and skill development for Early Career Scholars-Master students and young professional journalists at the university "Aleksander Xhuvani", Elbasan, Albania.

The product will be a working group research *Human-annotated dataset for social media sentiment analysis for the Albanian language* with the aim of publication in an academic journal.

Plans for future collaborations include continued engagement with the host institution and collaborating researchers to further develop and refine sentiment analysis models. The STSM has laid the groundwork for future research projects and collaborations within the Action network, fostering knowledge exchange and collaboration among researchers in the field of computational analysis of opinion.

In summary, the STSM successfully achieved its planned goals and made significant contributions to the Action's objectives, particularly in advancing research in computational analysis of opinion for the Albanian language. The outcomes of the STSM are expected to be disseminated through publications, and future collaborations will further enhance research in this area.