INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

PROJECT REPORT

Task for Course: DLBAIPNLP01 – Project: NLP

CONTENT

1. TASK

You can choose from the following tasks for your project report. Please choose one of them to work on in your project report.

Note on copyright and plagiarism:

Please take note that IU Internationale Hochschule GmbH holds the copyright to the examination tasks. We expressly object to the publication of tasks on third-party platforms. In the event of a violation, IU Internationale Hochschule is entitled to injunctive relief. We would like to point out that every submitted written assignment is checked using a plagiarism software. We therefore suggest not to share solutions under any circumstances, as this may give rise to the suspicion of plagiarism.

1.1 Task 1: Sentiment Analysis on Movie Reviews

Background: With the advent of social media, movie reviews have become a popular means of expressing opinions about films. As such, it is crucial for movie studios to be aware of the general public's sentiment towards their productions. Therefore, the task is to develop a natural language processing (NLP) system that can analyze movie reviews and determine the overall sentiment towards the movie.

Task: Develop an NLP system that can perform the following tasks:

- Collect movie reviews data from publicly available datasets or through web scraping.
- Preprocess the data by removing stop words, by stemming or lemmatizing, and encoding the text data for use in a machine learning model.
- Train a machine learning model using supervised learning algorithms such as Naive Bayes, Support Vector Machines, or Deep Learning, or any other suitable algorithm. The model should be trained on a labeled dataset that consists of positive and negative reviews.
- Evaluate the model on a separate test dataset to assess its accuracy and generalization performance.
- Once the model has been trained and evaluated, use it to analyze movie reviews and determine the overall sentiment towards the movie. The output of the system should be a binary classification indicating whether the sentiment towards the movie is positive or negative.

Evaluate each step of the proposed system in different scenarios, starting from a small dataset of movie reviews; then progressively test it on larger datasets. Use publicly available datasets (e.g., Stanford's Large Movie Review Dataset) or web scraping to collect data. Include a link to your GitHub or GitLab Repository with the code and report of the developed system, including the performance metrics on the test dataset.

Introductory literature:

Jurafsky, D. & Martin, J. (2013). *Speech and language processing [electronic resource]: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

Lane, H., Hapke, H. M., & Howard, C. (2019). *Natural language processing in action [electronic resource]: understanding, analyzing, and generating text with Python.* Manning.

Paramesha, K., Gururaj, H. L., Nayyar, A., & Ravishankar, K. C. (2023). Sentiment analysis on cross-domain textual data using classical and deep learning approaches. *Multimedia Tools and Applications: An International Journal*, 1–24.

1.2 Task 2: Text Classification for Spam Detection

Background: Spam messages are unsolicited messages sent in bulk to a large number of recipients. They can be a nuisance and can cause harm by delivering malware, phishing scams, and other fraudulent activities. Text classification is an NLP task that involves categorizing text data into predefined categories. In the context of spam detection, text classification can be used to distinguish between legitimate messages and spam.

Task: Develop an NLP system that can perform the following tasks:

- Collect a dataset of emails or text messages that are labeled as either legitimate or spam.
- Preprocess the data by removing stop words, by stemming or lemmatizing, and encoding the text data for use in a machine learning model.
- Train a machine learning model using supervised learning algorithms such as Naive Bayes, Support Vector Machines, or Deep Learning, or any other suitable algorithm. The model should be trained on a labeled dataset that consists of legitimate and spam messages.
- Evaluate the model on a separate test dataset to assess its accuracy and generalization performance.
- Once the model has been trained and evaluated, use it to classify new messages as legitimate or spam. The output of the system should be a binary classification indicating whether the message is legitimate or spam.

Evaluate each step of the proposed system in different scenarios, starting from a small dataset of emails or text messages; then progressively test it on larger datasets. Use publicly available datasets (e.g., University of California, Irvine's Spambase Data Set) or web scraping to collect data. Include a link to your GitHub or GitLab Repository with the code and report of the developed system, including the performance metrics on the test dataset.

Introductory literature:

Elakkiya, E., Selvakumar, S., & Leela Velusamy, R. (2021). TextSpamDetector: textual content based deep learning framework for social spam detection using conjoint attention mechanism. *Journal of Ambient Intelligence and Humanized Computing, 12*(10), 9287–9302.

Huan, H., Guo, Z., Cai, T., & He, Z. (2022). A text classification method based on a convolutional and bidirectional long short-term memory model. *Connection Science, 34*(1), 2108–2124.

Jurafsky, D. & Martin, J. (2013). *Speech and language processing [electronic resource]: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

1.3 Task 3: Text Classification for Topic Modeling

Background: Topic modeling is a technique used to identify the underlying themes or topics within a collection of texts. It has many applications, including market research, content analysis, and social media monitoring. Text classification is an important component of topic modeling, as it allows us to assign a document to a particular category or topic.

Task: Develop an NLP system that can perform the following tasks:

- Collect a dataset of annotated documents.
- Preprocess the data by removing stop words, by stemming or lemmatizing, and encoding the text data for use in a machine learning model.
- Train a machine learning model using supervised learning algorithms such as Naive Bayes, Support Vector Machines, or any other suitable algorithm. The model should be trained on a labeled dataset consisting of a corpus of text–tag sets.
- Evaluate the model on a separate test dataset to assess its accuracy and generalization performance.
- Once the model has been trained and evaluated, use it to classify documents. The output of the system should be a multi-class classification indicating the assigned category or topic.
- The model should be able to handle different types of texts, such as news articles, academic papers, and social media posts.

Data: There are many publicly available datasets of annotated documents that can be used for training and testing text classification models. For example, the Reuters Corpus and the 20 Newsgroups dataset are commonly used benchmark datasets for text classification. Additionally, domain-specific datasets can be used for topic modeling in particular fields, such as medical research or legal documents.

Evaluate each step of the proposed system in different scenarios, starting from a small dataset of the corpus; then progressively test it on larger datasets. Include a link to your GitHub or GitLab Repository with the code and report of the developed system, including the performance metrics on the test dataset.

Introductory literature:

Jurafsky, D. & Martin, J. (2013). *Speech and language processing [electronic resource]: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

Schröder, C., Müller, L., Niekler, A., & Potthast, M. (2021). *Small-Text: Active Learning for Text Classification in Python.*

Schulman, A. & Barbosa, S. (2018, December 12-14). *Text Genre Classification Using only Parts of Speech* [Conference presentation]. 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, Nevada, USA.

2. ADDITIONAL INFORMATION FOR THE EVALUATION OF THE PROJECT REPORT

When conceptualizing and writing the project report, the evaluation criteria and explanations given in the writing guidelines should be considered.

3. TUTORIAL SUPPORT

In this project report task, several support channels are open; as the student, it is your responsibility to select your preferred support channel. The tutor is available for technical consultations and for formal and general questions regarding the procedure for processing the project report. However, the tutor is not required to approve outlines or parts of texts and drafts. Independent preparation is part of the examination work and is included in the overall evaluation. However, general editing tips and instructions are given in order to help you get started with the project report.