Assessment 1

Identification of Toxic, Engaging, and Fact-Claiming Comments

Submission Instructions:

1. Upload your solution files to Canvas by **Sunday 21st of November 2021**.
2. Go to the Project_1 folder in Canvas to upload your files.
3. Once you have submitted your file, you should verify that you have correctly uploaded it. It is your responsibility to make sure you upload the correct file.
4. Please make sure you **fully comment on your code** that will reflect your **OWN** work.
5. You need to use **Jupyter Notebook** and run the code before the submission, and you can submit a .ipynb, .pdf and/or .html files.
6. Please put **your student name and number** as comments **at the top of your file**.

This project is worth 50% of the Natural Language Processing Module.

Motivation:

As the world is getting smaller and smaller with the proliferation of hand-held devices, an exponential increase in social media is also witnessed by the people living in the current era. On the one hand, social media can communicate brighter ideas exchange thoughts for the greater good, but on the other hand, it is also criticized for the spread of fake news and extreme hatred or toxic comment. In this project, we will be focused on the identification of toxic comments. Also, we will be extending the focus to two other classes of highly relevant comments. To moderators and community managers on online discussion platforms: engaging comments and fact-claiming comments, meaning comments that should be considered a priority for fact-checking.

Dataset:

For this we will be utilizing an annotated dataset of over 3,000 Facebook user comments that have been labeled by four trained annotators. The dataset is drawn from the Facebook page of a political talk show of a German television broadcaster, including user discussions from February till July 2019. The dataset is provided in anonymized form. User information and comment IDs will not be shared. Links to users are replaced by @USER. Links to the show are replaced by @MEDIUM, and links zu the moderator of the show are replacd by @MODERATOR. For trial data, a sample of user comments to two further shows has been provided. The user comments in the test data were drawn from discussions on different shows than in the training data. This way, we could provide a realistic use case and further could control a possible bias causes by topics. The annotation guidelines for the data set can be obtained upon request. The data is provided in .csv-format and the following structure:

comment_id	comment_text	Sub1_Toxic	Sub2_Engaging	Sub3_FactClaiming
1	*"Kinder werden...."*	0	0	1
2	*"Die aktuelle Situation zeigt vor allem..."*	0	1	0
...	...	...	...	...

Both the training and testing dataset will be provided in the Canvas by name of :

a. Assessment1 _Toxic_Train.csv and
b. Assessment1 _Toxic_Test_For_Evaluation.csv

Subtask 1: 😈 Toxic Comment Classification (Binary Classification Task)

The detection of **toxic content** in online discussions **remains challenging** and new approaches are constantly being demanded and developed.

message	Sub1_Toxic
"Na, welchem tech riesen hat er seine Eier verkauft..?"	1
"Ich macht mich wütend, dass niemand den Schülerinnen Gehör schenkt"	0

Subtask 2: 🤗 Engaging Comment Classification (Binary Classification Task)

In addition to the detection of toxic language, community managers and moderators increasingly express interest in **identifying particularly valuable** user content, for example, to highlight them and to give them more visibility. That includes rational, respectful, and reciprocal comments that can **encourage** readers to join the discussion, **increase positive perceptions** of discussion providers, and can enhance more **fruitful** and less violent exchange.

message	Sub2_Engaging
"Wie wär's mit einer Kostenteilung. Schließlich haben beide Parteien (Verkäufer und Käufer) etwas von der Tätigkeit des Maklers. Gilt gleichermassen für Vermietungen. Die Kosten werden so oder soweit verrechnet, eine Kostenreduktion ist somit nicht zu erwarten."	1
"Die aktuelle Situation zeigt vor allem eines: viele Kinder mussten erkennen, dass ihre Mutter bestenfalls das Niveau Grundschule, Klasse 3 haben."	0

Subtask 3: ☝ Fact-Claiming Comment Classification (Binary Classification Task)

Beyond the challenge to ensure non-hostile debates, platforms and moderators are under pressure to act due to the rapid spread of **misinformation** and fake news. Platforms need to review and verify posted information to meet their **responsibility** as information providers and distributors. As a result, there is an increasing demand for systems that automatically identify comments that should be fact-checked manually. Note that this subtask is not about the fact-checking itself or the identification of fake news. However, the identification of fact-claiming comments is a **preprocessing step for manual fact-checking**.

message	Sub3_FactClaiming
"Kinder werden nicht nur seltener krank, sie infizieren sich wohl auch seltener mit dem Coronavirus als ihre Eltern - das ist laut Ministerpräsident Winfried Kretschmann (Grüne) das Zwischenergebnis einer Untersuchung der Unikliniken Heidelberg, Freiburg und Tübingen."	1
"hmm...das kann ich jetzt nich nachvollziehen..."	0

What we are looking for :

a. Demonstrate your knowledge on the topic (problem and proposed solutions in last year only), shouldn't go beyond 4 pages of Word document with Times New Roman font and font size being 11. **[25% Marks]**
b. Define a strategy to accumulate more datasets on the same topic and formulate a Global dataset. **[10% Marks]**
c. Explore data augmentation techniques that can be applied to the dataset. **[6% Marks]**
d. Drafting the preprocessing pipeline for the global accumulated dataset. **[6% Marks]**
e. Building of Language model on the given and curated dataset for each of the three subtasks. **[24% Marks]**
f. Development of the machine learning classifier for task-based classification for each of the three subtasks. **[24% Marks]**
g. Use parsing technique along with the Language model for any of the task to build a classifier. **[5% Marks]**

Submission:

a. Submit all the code written in the process to achieve the task-based goals.
b. Submit a document that contains answers to all the questions asked in the previous section ('What we are looking for').

Code

a. All code should be completed using Python as the programming language.
b. Your code should have a logical structure and a high level of readability and clarity. Please comment on your code and put all code into functions. Your code should be efficient and should avoid duplication.

Late submissions

a. If you don't get the assignments done to your satisfaction and don't meet the minimum requirements by the deadline, you have the option (as with any assignment at CIT) of submitting up to 1 week late for a penalty of 10%.
b. This penalty is subtractive. Work that would have earned 55% if on time would get 45% (not 49.5%) if late.
c. The penalty is applied weekly. So, one day late costs the same as 6.

d. If you have a specific reason for submitting a late assignment (sickness, etc.), please submit directly a medical certificate to the department secretary.

Plagiarism

Please read and strictly adhere to the CIT Honesty, Plagiarism and Infringements Policy Related to Examinations and Assessments. Note that reports are checked against each other and external web sources for plagiarism. Any suspected plagiarism will be treated seriously and may result in penalties and a zero grade.

Grading

The assignment is worth 50% of the overall mark for the module. Marks will be awarded based on the quality of the code and the results. In particular, I will be checking to see if you are handling and preprocessing data correctly, carrying out exploratory analysis to gain insights, correctly performing model implementation, and critically documenting everything in a clear and concise way. The submitted code will also be checked to ensure that the work is your own.