Learning from Data – Final Project

General remarks

For the final project, we will look at the task of **offensive language identification.** Differently than in the other assignments, we ask you to be a bit creative: you should come up with your own research question. Grading will be determined by your final report. See the final page for the specific grading sheet. Also, please check out the slides of week 7, as it will contain important information. **This assignment is to be done individually.**

Deadline for submission on Brightspace: Monday November 4th, 2021, 23:59.

What you have to hand in by the deadline:

- The final research report. See Section 3 for details. Please, make sure to hand in a **pdf** file following the usual report template.

1 Data: Offensive Language Detection

The data comes from a *shared task*, which is a competition between researchers to create the best model for a certain task. The task concerned determining whether a given tweet contains offensive information or not.

Size and collection You will use the data that was used in the Shared Task on Offensive Language Identification in 2019. There are a few important papers to check out. First, the paper in which the data set is described:

- Predicting the Type and Target of Offensive Posts in Social Media. Zampieri et. al (2019)

This paper described in detail how the data was collected and annotated. Please check it out and make sure you understand the details.

Then, there is the overview paper of the shared task itself. This describes the results of the shared task and the best working system. Since it was 2019, the best performing systems used BERT, but not newer models. Note that a lot of teams participated, so there are many papers to check out!

- SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). Zampieri et. al (2019)

You will notice that the authors also provided annotated data on two subtasks. You do **not** have to work with that. We will only focus on the first task: determining whether tweets contain offensive information or not.

I have already processed the data for you. Please download the train, dev and test set from Brightspace. I made sure the test set is the exact same as in the shared task, meaning your F-scores are comparable to participants of the shared task. You should use this information in your report, of course.

Please check out the data manually to see what it looks like. You will notice it is very messy data (from Twitter) and that certain preprocessing steps are already applied. Perhaps this influences your future modelling decisions.

2 Research goal of the project

I want you to address two different goals in this research project.

- Creating the best possible model given modern enconder-only or seq2seq LMs
- Answering an additional original and creative research question

For the first goal, you are going to fine-tune language models. The second one is more interesting. You are expected to come up with an approach or research question that does a bit more than simply fine-tuning a language model to get a high score. You are free to choose your own research question. To give you some ideas:

- Looking into different preprocessing methods of the text, especially in relation the text being tweets (hashtags, emojis)
- Looking at the performance of mono-lingual language models of non-English data. Does this work at all? Does this tell us something about the languages?

- Extensive evaluation of additional features using an SVM. What type of features do you expect to help in this task? Did it work as expected?
- Using lists of offensive words as features. How well does a baseline model of only these features do? Are there obvious offensive words missing?
- Filtering offensive words from tweets to make the task harder. But does the task then still make sense?
- Similar as above, but filtering the best X features according to an SVM model
- Similar as above, but changing offensive words to a single OFFENSIVE token. Does the content of the offensive words matter at all?
- Using extra training data from related NLP tasks (e.g. hate speech detection). Does this work at all?
- Using automatically labelled data from OffenseEval 2020. Does this help? How much is needed? Can we use data from different languages?
- Checking how LMs deal with (artificial) noise inserted in the text. Can they still do the task?

Of course, whatever question you choose, you should check if there is any previous work that did something similar. Especially on the shared task itself people tried many different things already.

3 What you have to do

This section contains what you have to do for the project. Please read it carefully! Also do not forget to check out the **slides for week 7**, as it contains extra information. For example, you are expected to adhere to the best practices outlined in the lecture.

Models Independent of your research questions, we ask you to **at least** implement and provide scores for these four models on the given task:

- A baseline classic model using bag-of-words (e.g. Naive Bayes, SVM)
- A classic model with optimized feature set (e.g. POS-tags, character n-grams, etc)
- An optimized LSTM model with pretrained static embeddings (e.g. GloVe, FastText)
- A number of fine-tuned pretrained language models (i.e. BERT, RoBERTa, DeBERTa)

Luckily you more or less did all of this already in the assignments! Make sure to provide a description of these models in your report. Of course, you are free to use any other model you want! Please experiment!

Report You are also asked to produce a **report**. Again, you have to write this report as a **research paper.** The report should start with clearly explaining the problem you are working on, motivating why this is an important problem to work on. The research questions should be clearly laid out and you should cite previous work. Since this was a very popular shared task, there is a lot of previous work you can check out.

The method should contain the explanation of how you tackled this problem, a description of the algorithm(s) you chose to train your model, including parameter tuning and settings, any additional data/resources you incorporated, and how well you do when developing. You should also justify your choices explaining why you selected a certain approach, certain features, the learning algorithm, and so on.

In the results section you should clearly present all relevant results. How well do your models do compared to the best scores in the shared task? Do your models perform better, and if so, why? You should interpret the results if possible: how well are we actually doing on this task? Also include a section in which you discuss the results, with possibly an extra (error) analysis. End the paper with a clear conclusion and possible suggestions for future work.

Code There is no need to submit your code for the final project. However, you could potentially add a link to a public GitHub repository that contains the code for the project. This is good practice, anyway.

Grading sheet See below for the full grading sheet. You will get a grade for each criterion: the final grade is the average of all these grades. So make sure to check it out! And again, don't forget to check out the slides of week 7 for advice.

Writing	ATROCIOUS–BAD (1–4)	INSUFFICIENT (5)	SUFFICIENT (6)	GOOD (7–8)	EXCELLENT (9–10)
Structure	lacks a clear structure	inconsistent and/or not in agreement with conventions	logically structured in conventional sections	completely in line with contents and follows conventions where necessary	excellent (suitable for publication in a scholarly journal)
Coherence & Language	incoherent and difficult to follow, many grammatical errors or typos	parts of the report are not consistently linked at the level of chapters, sections, and paragraphs, report is hard to follow, English should be improved	parts of the report are sufficiently linked at the level of chapters, sections, and paragraphs	strongly cohesive at all levels (chapters, sections, paragraphs), level of English is good	extremely cohesive at all levels; professional argumentation throughout, excellent language use
Introduction	ATROCIOUS–BAD (1–4)	INSUFFICIENT (5)	SUFFICIENT (6)	GOOD (7–8)	EXCELLENT (9–10)
Motivation & Previous work	the rationale is missing and the study is not contextualised	the rationale is unclear and contextualisation is weak	the rationale is explained sufficiently clearly and the contextualisation is sufficient	the rationale is convincing and clear, the study is well contextualised	the rationale clearly shows that the study is important and the study is very well contextualised
Research questions	there are no research questions	the research questions are unclear, too general, or not linked to topic	sufficiently clear and feasible research questions linked to topic	clear and interesting research question that follows from the data set	very clear and interesting and challenging research question that follows logically from the data set
Method	ATROCIOUS–BAD (1–4)	INSUFFICIENT (5)	SUFFICIENT (6)	GOOD (7–8)	EXCELLENT (9–10)
Description	severely lacking description of used data sets and algorithm	barely any description of used data sets and algorithms	sufficient description of data sets and algorithms but certain parts are left out	good description of data sets and algorithms used	excellent description of data sets and algorithms
Reproducbility	method is lacking import details, reproduction impossible	method is lacking certain details, which makes reproduction hard	method contains most details but can be improved	method contains all details for reproducibility but some parts are not clear or irrelevant	excellent description of the method containing all relevant details
Soundness	experimental setup not sound, results cannot be trusted	experimental setup contains certain mistakes	experiment setup is sufficient, but more experiments could have been done	good experimental set up	excellent sound experimental setup covering all factors
Models	Clearly did not meet minimum requirements in terms of different models	Not enough models have been implemented, baselines are lacking	Sufficient implementation of a number of algorithms, baselines are OK	Implementation of a number of different algorithms with different feature sets or architectures, good use of baselines	Excellent use of different algorithms and feature sets/architectures, use of correct and fair baselines
Results	ATROCIOUS–BAD (1–4)	INSUFFICIENT (5)	SUFFICIENT (6)	GOOD (7–8)	EXCELLENT (9–10)
Overview	no clear overview of all results	weak (not all relevant results are shown or discussed)	sufficient overview of results, some parts still unclear or missing	good overview of all relevant results	excellent overview of results, showing all results in a logical manner
Performance	bad results (below baseline) and no explanation provided	weak results, no explanation provided	bad results with partial explanation provided, or decent results without explanation	good results with good explanation provided	excellent results, models are pushed to the best performance
Discussion & Analysis	lacking	weak (no interpretation of results)	sufficient discussion or interpretation of results, some extra analysis is performed	critical discussion of results, which are interpreted to give insights. Some nice extra analysis is performed	insightful, creative and critical discussion of results, great extra (error) analysis performed