Project Overview

1. Frame your task in a brief paragraph!

*Our task is to filter spam effectively in the Hebrew language.  
We'll explore whether the best ways that were found to filter spam in English will achieve the same (or higher) results in Hebrew.*

1. Name your project proposal with a short name that can be used as a title and in any communication about your project, at least temporarily

*HebSpam.*

Input and Output for the Core of The Project

In many ways, an NLP project is seen as a pipeline, one that needs to start from data and pass it through possibly several layers of processing. For example, one kind of pipeline can be: tokenization ⇢ morphological segmentation ⇢ part of speech tagging ⇢ parsing ⇢ semantic parsing). This is just one kind of flow between processing layers, others may omit/ unite some of the layers in this example or sport entirely different layers that make sense for the specific task at hand.

However, typically, in a research oriented project, only one key part of the pipeline is really the core of a well-framed research project. For example, you may need to segment a sentence into tokens/morphemes before e.g. parsing it in some novel way. In that example case, the segmentation layer/task is not really a core aspect of your project, if your main area of novelty is within the parsing layer.

Of course there can be many other examples for this. Below, we ask you to frame your core layer, so that you can then rigorously define the input and output to that layer. We may also call it your Core Task.

1. According to the above description, what is the Core Task of your project?

*The Core Task of our project is to find the most accurate classification of the Hebrew spam message by implementing different classification methods.*

According to the above description, very rigorously, what is the Input Definition for your Core Task?

*Sentences that are tagged as 0 or 1 for no-spam or spam, with at least 5,000 sentences, matching to the English spam dataset size.*

How do you rigorously define the Expected Output of your Core Task?  
*Prediction for every sentence whether it is spam or not.*

Data Definition

Framing what is the core of your task, and the structure of the input and output for that core is essential. You have just done that. What about adequate training data? have you found any fully suitable data? do you need to clean or pre-process it in order for it to be useful? do you plan to annotate your data? Describe your training data below.

Be precise and inclusive, if necessary, provide a link to any document on Google Drive to which you grant access to view. To avoid ill-defined or impossible projects, formalize it as much as possible, and take a really good look at data you choose to use before filling here. If necessary, go back to the previous stage to make sure everything is logically fully aligned.

1. Description of Your Data  
   if applicable here, you may include hyperlinks, or a hyperlink to any table or document where you document your chosen data / datasets.

*We couldn't find any public dataset Hebrew spam, so we will create two Hebrew datasets, one that contains email and one that contains SMS messages.*

*We will export our personal Email inbox, SMS messages, and Whatsapp to have both spam and non-spam messages in the datasets, then we will classify them.*

*For achieving a large number of messages and a variety of spam types we will ask from our environment to send us the spam messages they get also.*

*We will have to lower case the words and we will use the transliteration table from Maman 13, remove pictures and HTML tags from emails, and convert specific words to some var:  
names will be converted to \_name\_ , emojis = \_emoji\_ ,  
links = \_link\_ , \n = \_newline\_ , email addresses = \_email\_ ,  
phone numbers = \_phone\_ .*

*We can extend the dataset by merging it with the yields of 5000 sentences from the "heb-ctrees" dataset from Maman 13 labeled as non-spam.*

*\*The purpose of creating two different datasets is to try our model for both cases, finding whether they are even different and learning about the disadvantages of some methods and improve the model by eventually combining the datasets together.*

1. How well did you already inspect your chosen datasets or data sources, for suitability for the task? Scale 1-10. 5
2. Do you intend to annotate data as part of the project? Yes/No
3. Do you find gaps between the data you have chosen and obtained and the definition of the input for your task? Yes.
4. Do you find gaps between the data you have chosen and obtained and the data that you think you need for your task? No.

Suggested Significance of the Task

1. How did you come up with this task framing?

*We are usually reading out emails and SMS messages and do not want to leave them unread. We often find ourselves opening unwanted messages and classifying the sender as a spammer rather than the message as spam.*

1. Why is this task personally interesting to you, if at all?

*Because people all over the country experience it every day with SMS messages and emails (mostly when elections are coming soon), so we would love to know how spam filtering works and fit it the best we can do for Hebrew.*

1. In what ways do you find the task valuable or interesting for a wide (or wider) audience?

*We live in a state where our communication is mostly in Hebrew, therefore our texts are mostly in Hebrew as well, and our spam also.*

*Therefore, we need to do our best efforts to prevent our inbox from filling up with spam in Hebrew (supposing filtering spam in English is satisfying).*

*Spam, except for being annoying, can harm people and cause problems such as engulfing important personal mail, wasting network bandwidth, consuming users’ time and energy to sort through it, etc.*

Challenge/Difficulty and Learning Potentials

1. What do you think makes it a challenging task for you to be working on?

*Because of some variance between languages, sometimes data processing in English doesn't fit to Hebrew. Therefore some steps must be fitted to the language characteristics.*

*It will be challenging to use and suit the known methods for spam detection in English to Hebrew, finding if they are good enough (By the articles it is summed up by using feature selection, SVM, AdaBoost, Maximum Entropy, Memory-Based Learning, Naïve Bayes).*

*With that baseline, we also want to explore whether methods other than the known ones may be more satisfying (such as finding if there are any major differences between parse trees of spam and non-spam messages ?).*How likely do you think you are to succeed in arriving at a worthy work, despite these challenges?  Scale 1-10 7

1. Which elements you have studied or experienced during the course (or in prior learnings), will come into play in your project as framed? (checkbox)  
   - Sequence Tagging  
   - Constituency Parsing  
   - Dependency Parsing  
   - Formal Semantics  
   - Morphological Analysis or Segmentation  
   - Language Modelling  
   - Other Graph Algorithms  
   - Classification  
   - Deep Learning  
   - Lexical Semantics  
   - Other \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. In what specific ways if any, do you currently expect that working on this project will expand your know-how or expertise? אין צורך למרוח

*It will implement our knowledge practically.*

Evaluation Metrics

1. Which known evaluation metrics should apply to your task? are they a perfect fit? why so?

*Precision, Recall and F1 Score.*

*Cost-Sensitive – classifying non-spam as spam may be worse than the opposite.*

1. Do you plan developing your own evaluation metrics? if so describe briefly the justification/ rationale and which known metrics may/should inspire you in this

*No.*

State of the Art and Prior Work

1. What are some key prior art publications applicable to your task as framed? provide pairs of hyperlinks and titles / short descriptions!
2. An article describing trying different methods for spam detection, quoted 383 times. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.7685&rep=rep1&type=pdf>
3. An article describing a set of experiments focused on the AdaBoost Algorithm, quoted 383 times.   
   <https://arxiv.org/pdf/cs/0109015.pdf>
4. A blog post containing a reference to an English spam dataset.  
   <https://towardsdatascience.com/spam-or-ham-introduction-to-natural-language-processing-part-2-a0093185aebd>
5. Provide a link to a google document or table where you detail the importance of each of the key publications you've identified, to your task.  
   <https://docs.google.com/document/d/1-h_p9LoiedkmxYPetcUluAlyMbAMeD1IYRiWKThrEs4/edit?usp=sharing>
6. What is the novelty of your anticipated outcome, relative to those prior art?

The novelty of our anticipated outcome is using different data from a different language trying to equalize or improve the accuracy for Hebrew.

Room for Evaluation and Analysis

1. What role should/would evaluation and/or results analysis play, in your project as scoped? Scale 1-10 8

Use of Existing Solutions, Libraries and Tools

1. Do you find good reason to leverage existing toolkits, libraries and tools for certain layers of the text processing pipeline necessary under your project framing, other than your own code? please suggest clear justification if so.

Yes, we will use existing libraries for some algorithms rather than implementing them on our own, for example, the sklearn library supplies a large set of tools like Naive Bayes and AdaBoost classifiers.  
We don't see the direct implementation of an existing algorithm as a mission of our project, but we will implement what's needed as long as we can't find the right response to our needs.

1. Please enumerate your anticipated technology stack, from programming languages to frameworks and cloud services.  
   Python, Pycharm.

Acknowledgement

1. We as a team, understand that all results claimed and reported in our final report, should be fully reproducible via our code submission, which should be hosted and submitted via a Github repository. Yes/No/Other
2. We understand that the evaluation criteria for our project are listed at [http://bit.ly/2XhTpqf](https://www.google.com/url?q=http://bit.ly/2XhTpqf&sa=D&ust=1562144016175000&usg=AFQjCNF6JePU3DzV6HfYfrZ4jM8Ucy8CaQ), and that we should engage the course team to clarify on any aspects thereof near the outset of the project. Yes/No/Other
3. We acknowledge that without careful final reviews of the applicable requirements represented via the link above, submitting our final report is likely to score very low on some key aspects, and that we should perform such internal (and possibly iterative cycles of) reviews within our team in order to fulfill the potential of our work. Yes/No

סטטוס ההצעה

* טיוטה
* מוכנה להתייחסות ראשונית של צוות הקורס
* מרגישים ממש מוכנים