Application for 2021 UiO:Energy Summer Research Project

Machine Learning Approach to Public Sentiment Analysis towards Wind Energy in Norway

Natalia Sirotko-Sibirskaya, Oskar Vågerö, Marianne Zeyringer

Project summary

1. **Name of supervisor(s):** Natalia Sirotko-Sibirskaya, Oskar Vågerö, Marianne Zeyringer

2. **Preferred background of candidate(s):** Based on the project description as below an ideal candidate should possess the following qualifications:

   (a) Be familiar/proficient with/in Python and possibly with social media data scrapping, or large datasets, in general,

   (b) Be familiar or willing to learn new machine learning methods relevant for natural language processing (NLP),

   (c) Be proficient in the Norwegian language enough to understand and manually classify social media posts including not immediately obvious ones such as sarcastic ones and so on.

   Since the project is of a limited duration (6 weeks) we think that an ideal candidate should be at least acquianted with programming in Python and be able to analyze data in Norwegian, whereas project-specific tasks such as particular NLP methods can be learned on the project given candidate's natural affinity and strong interest for quantitative/programming tasks. The candidates who want to do the project for 12 weeks on 50 % are welcomed as well.

3. **Number of available projects (one or two):** one or two

4. **Preferred project period:** between April and October 2021

Outline of project work including expected outcomes/deliverables

We propose a research project within one of the four main research areas of the UiO, namely, Energy Transition and Sustainable Societies. With renewable energies gaining more and more importance we need to understand better how sustainable energy solutions are being accepted by the society. A traditional way to collect public opinions is via questionnaires, surveys or interviews. However, these methods are often prone to a selection bias, missing data or incomplete information. One of the alternative resources which can be used in order to develop an understanding of how renewable energy and wind energy, in particular, is accepted in Norway is to analyze social media such as Twitter, Facebook, Instagram as well as publicly available commenting of the news media. Although being susceptible to certain biases such as, e. g., age bias since most Twitter users are under 60, see `https://www.statista.com/statistics/585035/twitter-users-in-norway-by-age-group/`, or a geographical bias, i. e. social media users might be more active in large cities compared to small neighbourhoods, the rich data environment of the contemporary world calls for exploration and systematization.

Our project can be best exemplified by the recent work by [Kim et al., 2020], where the authors analyze Twitter data for the U. S. and reconstruct a map of the U. S. in terms of public sentiments towards the solar energy. For the U. S. it is particularly meaningful to perform such public sentiment analysis using Twitter data since the number of Twitter users is close to 70 millions providing rich data for (almost) any socially relevant question, see `https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/` for Twitter statistics in the U. S. The main machining learning tool which the authors rely on in their work is developed by [Liu et al., 2019] called Robustly Optimized Bidirectional Encoder Representations from Transformers (ROBERTa). ROBERTa has an advantage of possessing an extensive built-in pre-training dictionary based on Wikipedia and other media so that it does not have to be trained excessively. Using Twitter Application Program Interface (API) in Python [Kim et al., 2020] filter out the tweets on the solar energy between January 2020 and early July 2020. As a result about 70 000 tweets relevant for the analysis are obtained. After manually processing around 5 000 tweets to fine-tune ROBERTa for a better classification between positive and negative attitude towards the solar energy [Kim et al., 2020] classify the rest of the relevant tweets and reconstruct spatiaI patterns in sentiment toward solar energy providing a quantitative score for each of the 51 U. S. states. In a nutshell, the goal of our project is to implement the idea as in [Kim et al., 2020], but for the case of public acceptance of the wind energy in Norway.

Compared to the example of U. S. data as above in our work we face the following challenges. First of all, the number of Twitter users in Norway is around 700 000, see [Bruns and Enli, 2018]. This is only about 5 % less than the percentage of the Twitter users in the U. S. when measured as a ratio of the total number of Twitter users to the current population, however, in absolute and not in relative numbers it might also mean that there is not enough data to extract public sentiments. Consequently, one might need to consider several sources such as, e. g., Facebook or publicly available commentaries of the relevant articles in the largest Norwegian newspapers online. Second, the NLP method such as ROBERTa cannot be used directly in the context of the Norwegian language for the obvious reason that the libraries are in English. Recently, a number of open-source Norwegian NLP tools have been developed such as a Nordic BERT, see `https://github.com/botxo/nordic_bert`, or so called spaCy, `https://github.com/web64/spacy-norwegian`, which can be used in the context of the suggested project. In particular, we are interested in testing in practice and employing a Norwegian BERT developed recently at the Department of Informatics of the University of Oslo, see `https://www.mn.uio.no/ifi/english/research/groups/ltg/news/20210113.html`. We would like to stress that the goal of our project is not to develop a novel methodology for the sentiment analysis of the Norwegian text, but rather to apply existing tools in such a way that we can reach the desired goal, however, depending on the qualification of the candidate we might allow for certain freedom given that the project is completed on time.

At each of the project's phase such as data scrapping from different social media, choosing an appropriate NLP algorithm, choosing a training set including manual data analysis, making test runs, validating the results the successful candidate is expected to participate actively either by following the recommendations of the supervisors and/or suggesting her/his own ideas. To our knowledge such public sentiment reconstruction based on social media has not been performed in the energy literature for Norway up to now and we believe that with our project we can provide a meaningful contribution to a better understanding of social resonance towards the wind energy. The project results should (ideally) become a part of a scientific publication with open-access codes and possibly other publicly available interactive resources.

References

[Bruns and Enli, 2018] Bruns, A. and Enli, G. (2018). The Norwegian Twittersphere: structure and dynamics. *Nordicom Review*, 39(1):129–148.

[Kim et al., 2020] Kim, S. Y., Ganesan, K., Dickens, P., and Panda, S. (2020). Public sentiment toward solar energy: Opinion mining of twitter using a transformer-based language model. *arXiv preprint arXiv:2007.13306*.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). ROBERTa: A robustly optimized BERT pretraining approach.