A new experimental quality of life statistic for Flanders using machine learning and sentiment analysis on Tweets

**Keywords:** experimental statistics, new data sources, sentiment analysis, social media, official statistics

# Introduction

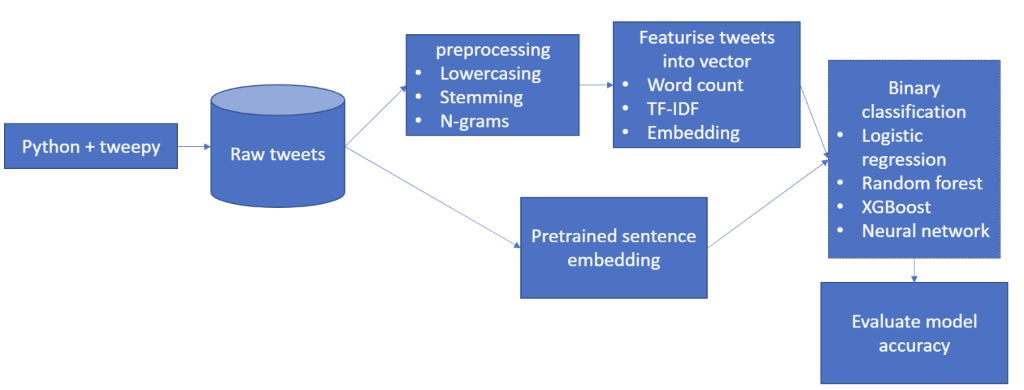
This paper presents the experiments and developments of Statistics Flanders on the creation of a new experimental quality of life statistic by using sentiment analysis on Flemish tweets. This study has been set up to test the use of big data sources like Twitter to measure the sentiment of a population (e.g. perceived quality of life, beyond GDP). Can the combined use of big data, text mining and machine learning provide a valuable alternative for surveys? The study provides a use case for Flemish-language tweets in Flanders (Belgium). Other cases for e.g. Poland (Polish) and Mexico (Spanish) are already available in the field of official statistics [1,2]. The case for Flanders has some special issues such as the relatively low penetration of Twitter in Belgium compared to other countries. The absence of a Flemish language-category in Twitter is another specificity. Even though the sentiment coding of tweets is the focus of this report, these specificities make the Flemish case interesting to develop further.

Current quality of life statistics are produced via surveys, which results in infrequently available statistics. Using social media data, daily (and even hourly) measures of quality of life statistics can be produced. This high frequency statistic allows for the immediate monitoring of the impact of events and interventions on the sentiment of the population.   
However, there are several challenges to be overcome in order to create such a statistic that is accurate, representative of the total population and interpretable.

Our study compares and evaluates different methodological choices that need to be made in the creation of such a statistic. We compare different methods for creating an annotated dataset, natural language pre-processing techniques and different machine learning algorithm choices for the sentiment model.

# Methods

A schematic overview of the sentiment analysis pipeline including considered alternative methodological choices can be seen below in Figure 1. The code for each step can be found on our GitHub repository [3].



**Figure 1. schematic overview of approach**

Before obtaining Twitter data, one has to register for a Twitter account, get API keys and create a new app on the Twitter website. Actual collection of the data is performed by calling the REST Twitter API from specific-written Python code. The second step is to normalize the text (removing stop words, lemmatization, removing special characters, etc..) and perform feature extraction (derivation of meaningful features from the raw data that can be used by a machine learning model to make predictions).

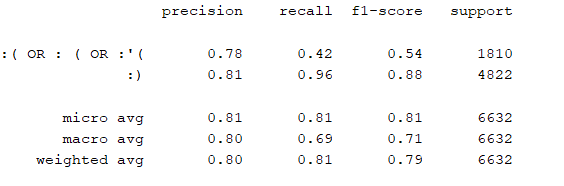
The machine learning part of the process starts with coding a training data set. This coded training set serves as examples of positive and negative tweets, from which the (supervised) machine learning program then tries to learn patterns. The result is a classification model. This model is expected to predict outcomes (sentiment) for new tweets in the future. In this paper we evaluate several machine learning models: penalized logistic regression, random forest, gradient boosting trees and several deep learning architectures.

An important aspect of the process is the evaluation of the model: how well are new tweets predicted by the model? In this paper we use different evaluation measures (accuracy, recall, precision, etc.) and apply them to different machine learning algorithms to compare the performance of these algorithms. In our experiments we use a 5-fold cross validation approach on a dataset of 19.000 positive and 7.000 negative tweets.

# Results

Table 1 contains the evaluation report of the count vectorization – logistic regression model with 81% accuracy. Other models achieved similar performance but do not exceed the performance of this relatively simple vectorization scheme and classification algorithm. This evaluation report gives us some interesting insights besides pure accuracy. The precision and recall metrics show us that the model performs better on positive tweets than on negative tweets. This could be because of the larger training set available for positive tweets.

**Table 1 : Performance of count-vector + penalised logistic regression on test set**



Following the promising results of these experiments, we have developed a completely automated cloud-based pipeline that retrieves Flemish tweets, automatically annotates their sentiment and generates an aggregated experimental statistic in a simple user interface.

# Conclusions

We believe that the technical challenges overcome in our work will be of great interest to the statistical community. We provide solutions to several technical challenges, and by consistently producing this experimental statistic, we have built an environment for further experimentation.

It is interesting to note that new experimentation is already underway to include state of the art language models (such as GPT-3 [4]). The results reported in Section 3. already show that strong sentiment classification performance can be achieved, however, research has shown that these new language models often provide significant performance benefits in natural language processing tasks [4]. It will be interesting to see if we can observe similar model improvements by including these models in our pipeline.

By experimenting with and developing this experimental statistic we have answered many methodological questions that bring us closer to the development of a useful statistic based on the sentiment of Flemish tweets. However, we have also identified several research questions that still need answering before we feel confident publishing this statistic as an official statistical institute.  
Our main area of further research is the representativeness of Tweets. As there are significant biases in the set of Twitter users, it cannot be considered as a representative sample of the Flemish population. More work is needed to better describe these biases and to alleviate their impact on the resulting statistic. Furthermore, we would also like to research the difference between a person’s sentiment and the sentiment of their tweets: what are the differences between the sentiment of the Flemish Twitter-sphere and the general sentiment in Flanders?  
We are also looking into expanding our concept of sentiment, currently coded as positive or negative, to include more fine-grained emotions.  
Finally, we are also investigating the potential to look at the sentiment in tweets on specific topics, such as the sentiment on COVID-19 measures, sentiment on lifelong learning, etc.

# References

1. GitHub. 2020. *Jmaslankowski/WP7-Population-Life-Satisfaction*. [online] Available at: <https://github.com/jmaslankowski/WP7-Population-Life-Satisfaction> [Accessed 15 October 2020].
2. (INEGI), I., 2020. *Estado De Ánimo De Los Tuitero En México*. [online] Site Name, i.e. Moz. Available at: <https://www.inegi.org.mx/app/animotuitero/> [Accessed 15 October 2020].
3. Reusens, M., Callens, M., 2020. *Mireusen/Hlmos-Statistiek-Vlaanderen-Twitter*. [online] GitHub. Available at: <https://github.com/mireusen/hlmos-statistiek-vlaanderen-twitter> [Accessed 15 October 2020].
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.