Data Science for Transportation (2024) 6:5
https://doi.org/10.1007/s42421-024-00090-5

RESEARCH

Performing Sentiment Analysis Using Natural Language Models for Urban Policymaking: An analysis of Twitter Data in Brussels

Floriano Tori[1] · **Sara Tori**[2] · **Imre Keseru**[2] · **Vincent Ginis**[1,3,4]

Received: 18 July 2023 / Revised: 8 November 2023 / Accepted: 28 February 2024 / Published online: 11 April 2024
© The Author(s) 2024

Abstract

Objectives Mobility is a core challenge to transition towards sustainability. Cities are, therefore, rethinking their mobility to reduce negative externalities such as (greenhouse) gas emissions or congestion. When trying to implement sustainable urban mobility plans, there is often resistance from citizens. This can indicate a disconnect between the public and policymakers due to a lack of participation, coupled with the fact that current data-collection methods often used (such as travel surveys) are limited in scope. Advances in big data analysis and user-generated content provide opportunities to gain deeper insights into citizens' perceptions of mobility policy changes. This paper explores how sentiment analysis through deep learning can be used in transport planning.

Materials and Methods In this research, we analyse the sentiments of citizens towards recent changes in mobility policy in Brussels, Belgium, through Twitter data. We analyse 1998 tweets about changing mobility policy in Brussels between July 18th, 2019 (forming of the last Brussels regional government), and December 31st, 2022 (starting date of the analyses). For our analysis, we employ two pre-trained language models: *XLM-T* and *GPT4*.

Results Our results show that the sentiment with regard to the new mobility interventions is, as reflected by Twitter posts, not overwhelmingly negative. Furthermore, we find that the performance scores of *XLM-T* when no domain-specific fine-tuning has occurred (zero-shot evaluation) is fairly low (0.48). Once the model is trained on our domain-specific data, it reaches an accuracy of 0.67. When using *GPT4*, the model reaches an accuracy of 0.66. Additionally, *GPT4* seems better suited at identifying mismatched tweets, i.e. tweets using vocabulary that has a different sentiment than the one the tweets expresses (e.g., sarcasm). This might indicate that large language models might be better suited to obtain implicit sentiments expressed in a text.

Conclusions From a machine learning perspective, our experiments highlight the difficulty of recognising contextual sentiment (in this case, a sentiment towards changes in mobility policies), which may differ from the sentiment reflected in the vocabulary used. This is especially important if these two sentiments do not correspond, a problem both models struggled with. Additionally, we show that *GPT4* can provide additional information when performing sentiment analysis by prompting it to attribute scores to texts. This paper opens new perspectives on understanding and addressing public sentiment in urban mobility policies. The advancements in language models, and the effective integration of user-generated content, can provide policymakers with a more comprehensive understanding of public sentiment, facilitating the implementation of certain policies.

Clinical Relevance None.

Keywords Sentiment analysis · User-generated content · Urban mobility · Large language models · Transport planning

Introduction

Cities worldwide are grappling with a core challenge: the transition towards more sustainable mobility. Current car-centred transport is linked to a number of negative externalities, including greenhouse gas emissions, air pollution, congestion, accidents, and noise pollution (Parry et al. 2007;

Floriano Tori and Sara Tori have contributed equally and share first authorship.

Extended author information available on the last page of the article

Small and Verhoef 2007). In Europe, transport is responsible for 27% of greenhouse gas emissions, with road transport representing the greatest share of these emissions (72% in 2019) (European Environment Agency 2022). Cities are therefore rethinking their urban mobility systems, with car use declining in advanced cities that promote sustainable alternatives to the car (Jones 2014). However, we often see that the transformation process towards sustainable mobility is met with protests. This was the case in Barcelona, where the superblock model aims to decrease through-traffic (O'Sullivan 2017), or for the expansion of London's congestion charge (Milmo 2007). More recently, this has also been the case in Brussels, Belgium, where the phased implementation of several low-traffic neighbourhoods (LTNs) resulted in sometimes violent protests (The Brussels Times 2022).

Although the preconditions for a sustainable mobility transition are known and well documented, mobility behaviour is still dominated by car use (Haustein and Kroesen 2022), with attempts at transitioning towards sustainable mobility often undermined by public resistance. These demonstrations are amplified by media coverage, raising awareness of the issues at hand (Jennings and Saunders 2019). One possible explanation for this is that forcing a system to change can create a backlash (Rotmans et al. 2012). Zipori and Cohen (2015), e.g., mention that it is necessary to implement changes in 'gentle' ways to avoid resistance, but that such an approach has its limitations.

When looking at the data collection for mobility planning, it is often still dominated by traditional methods, such as surveys, to estimate travel demand and transport supply. These methods are expensive and time consuming (Zannat and Choudhury 2019). There are therefore new demands being placed on data in terms of the amount of data required, as well as the accuracy and completeness of it (Stopher and Greaves 2007). Additionally, public participation and involvement have become core elements in transport planning, but can be challenging to achieve (Evans-Cowley and Griffin 2012). Developments in big data analysis can provide opportunities for mobility planning by complementing these traditional methods (Pucci and Vecchio 2019). One advantage of big data is that the sample size analysed can be larger than with traditional survey methods.

A particularly interesting direction to gain the necessary insights is through user-generated content (UGC), which can be complementary to traditional data-collection methods (Martin-Domingo et al. 2019). UGC is content that can be made widely available by individuals without needing to go through a publisher. Through social media, this has become possible for almost anyone (Wyrwoll 2014). UGC can be analysed using text mining techniques, one of which being sentiment analysis, where positive or negative opinions about a subject are analysed (Quan and Ren 2016). Using social media data is a relatively new development in transport planning, but it shows great potential (Nikolaidou and Papaioannou 2018).

In this paper, we therefore ask the following question: "Can sentiment analysis through pre-trained language models improve our understanding of public perception of mobility measures and interventions?" This paper seeks to provide policymakers and practitioners with an understanding of alternative tools for mobility planning that provide a broader understanding of public sentiment. Our analysis focuses on Brussels, Belgium, where the recent implementation of the regional mobility plan rerouting and restricting car traffic led to opposition in several neighbourhoods. We perform sentiment analysis using two different Transformer-based pre-trained language models: XLM-T (Barbieri et al. 2022), an encoder-based model fine-tuned on collected data, and GPT3.5/4 (OpenAI 2023), a decoder-based model employed in a zero-shot manner.

This paper is structured as follows: the next section provides some background on UGC and sentiment analysis. In the subsequent section, we introduce the mobility interventions that served as the subject of our sentiment analysis in Brussels, Belgium, and we explain the methodology employed. The penultimate section presents our results, and the last section provides a discussion and some concluding remarks.

Literature Review

User-Generated Content in Transport

Understanding the sentiments of the public can be a difficult task. In recent years, UGC started playing an important role across politics, business, and entertainment (Gal-Tzur et al. 2014). The availability of UGC allows for sentiment analysis in different areas, such as the harvesting and analysis of opinions and product trends (Tuarob and Tucker 2015), or political orientations (Maynard and Funk 2012).

In transport planning, travel surveys have historically been used to collect data and guide decision making. Surveys are useful to obtain socio-demographic information, but are labour-intensive and therefore more costly. These higher costs lead to smaller sample sizes, as well as lower update frequency of the data. Additionally, quality issues of the data can arise (Serna et al. 2017; Zannat and Choudhury 2019). Transport planning also faces difficulties with regards to public participation (Evans-Cowley and Griffin 2012), which is a necessary precondition for achieving sustainable mobility (Lindenau and Böhler-Baedeker 2014). UGC can provide an interesting complement to the traditional data-collection methods currently used in transport planning, as this type of data contains a high level of accuracy at a lower cost

(Zannat and Choudhury 2019). As such, UGC has been used to analyse the experience of transportation services (Collins et al. 2013), or the reporting of heavy traffic (Endarnoto et al. 2011). Serna et al. (2017) employ UGC to identify sustainability issues related to urban mobility.

Yet the full potential of UGC for the transport sector has not yet been reached (Gal-Tzur et al. 2014), and planners should further develop the use of social media as a data source (Lock & Pettit 2020). According to Kuflik et al. (2017), UGC has the potential to complement, enrich, or even replace traditional data collection in the transport sector. The integration of big data in the planning process can help reduce the duration of the planning cycle (currently ranging anywhere from 5 to 20 years (Khan et al. 2014)), as well as result in more informed and agile decision-making (Semanjski et al. 2016). The use of UGC to improve transport decision making by understanding the public's feeling towards mobility policies therefore offers an interesting avenue of research.

Sentiment Analysis on UGC

Sentiment analysis is a natural language task that analyses individuals' opinions, attitudes and emotions towards entities such as products, services, organisations, locations and events (Liu 2015). Sentiment analysis can encompass many approaches. In this work, we focus on simply classifying the polarity (i.e., positive, neutral, or negative) of text. It should be noted that there appears to be a 'negativity bias' within UGC, with social media being a sharing arena that reflects negative emotions (Jalonen 2014).

Various domains have successfully applied sentiment analysis to Twitter[1] data, e.g., from understanding the public's sentiment towards the COVID-19 pandemic (Naseem et al. 2021), to extracting trends of food consumption across the United States (Widener and Li 2014), or even predicting stock market evolutions (Pagolu et al. 2016). In the transport sector, Twitter-based sentiment analysis was used to evaluate the satisfaction of transit service users in Los Angeles (Luong and Houston 2015) and Chicago (Collins et al. 2013). Collins et al. (2013) find that users are more likely to express a negative sentiment. Lock and Pettit (2020) use Twitter data to evaluate public transport performance in Sydney, Australia, and found that there was no clear majority of either positive or negative sentiments being expressed. However, they also report that sarcasm was often not picked up and that sarcastic tweets were often labelled as positive. In their study, they compare two different models to perform the sentiment analysis, and conclude that the use of multiple models adds confidence to the interpretation of their results.

Sentiment Analysis Using Pre-trained Language Models

The Transformer architecture (Vaswani et al. 2017) has been an incredible advance for the application of deep learning on Nature Language Processing (NLP) tasks. Pre-trained Language Models (LMs) based on the transformer architecture such as OpenAI's GPT4 (OpenAI 2023), Google's PaLM 2 (Google, 2023), BERT (Devlin et al. 2019) or RoBERTa (Liu et al. 2019) are trained to create contextual word embeddings with the use of large amounts of unlabelled training data. Once pre-trained, these models can be fine-tuned for specific NLP tasks, which can be mono- or multilingual. State-of-the-art performances on multilingual tasks have been pushed by pre-trained multilingual models such as mBert (Devlin et al. 2019), XLM (Lample and Conneau 2019) or XLM-R(oBERTa) (Conneau et al. 2020).

Using social media data, specifically Twitter data, for NLP tasks does suffer from drawbacks due to its uncurated nature (Derczynski et al. 2013). Tweet brevity incentivises users to compress their message, omitting possible contextualising words (Derczynski et al. 2013). Additionally, the widespread use of slang and neologisms means Twitter data contain peculiarities which are generally not included in the general training corpus of language models (Camacho-Collados et al. 2020). Emojis also play an essential role in understanding social media data as they carry a non-negligible semantic load (Barbieri et al. 2018) and are omnipresent in their usage (Barbieri et al. 2017). This means that an NLP task such as sentiment analysis needs to consider this additional source of information when making predictions. Felbo et al. (2017) showed that training models on emoji prediction tasks improved their performance on other tasks such as sentiment analysis or sarcasm detection.

Although LMs are pre-trained on a large corpus of data, our topic-specific task poses some challenges. Judging the sentiment of a tweet requires knowledge about the domain of the subject, which is why models are further fine-tuned for a specific task and topic. Improving the performances can occur in different ways. A first approach consists of continuing pre-training. Gururangan et al. (2020) show that further pre-training on domain-specific and task-specific data offers performance gains. In the same direction, Rietzler et al. (2019) demonstrate that the performance of BERT for Aspect-Task Sentiment Classification (ATSC), which combines both aspect extraction and sentiment polarity detection, can be improved by further pre-training on domain-specific data and then fine-tuning the model on task-specific data. This combination of task and domain knowledge enhancement also works when using domain knowledge for

[1] Our research was performed before Twitter was changed to 'X'.

further pre-training but fine-tuning on (out-of-domain) task data (Xu et al. 2019), where this was demonstrated using BERT on tasks such as Review Reading Comprehension, Aspect Extraction and Sentiment Analysis. Further training can also be done purely through fine-tuning, whether on a task level (e.g. Glue benchmark tasks (Liu et al. 2019), intent detection/classification (Zhang et al. 2021) or text classification (Howard and Ruder 2018)), or on domain-specific level (Araci 2019).

Enabling a modal shift is dependent on changes in policy and planning. However, the developments described above demonstrate that traditional data collection methods alone are no longer sufficient for transport planning. Developments in big data, and UGC, can be a valuable complement to inform decision makers on the feelings of the public, but there is a need to understand how these methods can be useful to decision makers.

Materials and Methods

Research Context

Brussels, Belgium, is home to 1.2 million inhabitants (IBSA 2022). It is historically a very car-oriented city. The World Expo of 1958 provided a push to modernize the city, resulting in modern road infrastructure to accommodate cars (Hubert 2008). In recent years, there has been a trend to reclaim some of the urban space, with the most notable project being the conversion of one of the city's central car arteries into a pedestrian area in 2015 (Hubert et al. 2017). The city also adopted Good Move, its regional mobility plan, in 2020, which was the result of a four-year participatory process. The plan won the 2020 SUMP Award for its ambitiousness (Bruxelles Mobilité, 2020b). It includes the implementation of one of Europe's largest 30 km/h-zones, as well as the elimination of through-traffic in multiple neighbourhoods (Bruxelles Mobilité, 2020a). In the context of the COVID-19 pandemic, which hit Europe and Belgium in the spring of 2020, changes with regards to urban mobility in the city were also accelerated, with 40 new bike lanes being deployed faster than anticipated, or the closure of streets to cars (see for example Bruzz 2020a). However, the temporary closures to cars during COVID-19 sparked some backlash (Bruzz 2020b; Macharis et al. 2021), as did the phased implementation of the LTNs planned in the context of the Good Move plan (The Brussels Times 2022). These last protests resulted in multiple municipalities delaying or cancelling the implementation of their LTN. However, it is not actually clear how many people objected or supported the plans, since no surveys were carried out.

Methodological Approach

There are two main steps to approach our task of sentiment analysis for UGC data in the context of mobility changes in Brussels. First, we comb through Twitter to obtain relevant tweets. These are tweets which have as subject one of the (future) mobility interventions in Brussels, included in the regional mobility plan Good Move). In Northern Europe, 81% of the population is an active social media user (i.e., a user logging in in a 30-day period) (We Are Social & Meltwater 2023). When compared to survey data, UGC can provide complementary, faster, and specific information about a topic (Endarnoto et al. 2011). Using Twitter specifically as a source for sentiment analysis can be motivated due to multiple aspects. Sentiment analysis through Twitter is more straightforward than other platforms based on the post lengths, which are limited to an upper maximum[2] (Nikolaidou and Papaioannou 2018). Other platforms such as Instagram, Flickr or Foursquare do not offer the possibility of opinion sharing. Facebook allows users to share their opinions, but the data on Facebook for sentiment analysis has been found to be messy and not structured properly. Due to its low cost, ease of access and presence of (public) leaders (Naseem et al. 2021), Twitter offers an approachable way to voice concerns or praise about policies. Twitter data also offer spatiotemporal information about users sharing their opinion, as tweets are tagged with their time of posting and possibly a geotag, offering additional dimensions for analysis. Lastly, Twitter is the sixth most visited website worldwide (We Are Social & Meltwater 2023).

Once we collect the tweets, we use two pre-trained language models (*XLM-T* and *GPT4*) to analyse the sentiment of those tweets. When processing textual data, LLMs are used to create contextual numerical representations, i.e. embedding vectors, of the sentences using the transformer architecture (Vaswani et al. 2017). This embedding can then be used in different ways, depending on the model architecture (see Fig. 1; Sect. 3.2.2 provides more details on our specific use of the models).

For *XLM-T,* an encoder-based model, tweets are passed as input and the model outputs a probability for the three possible sentiments of the input. The one with highest probability is then chosen as final sentiment. In the case of the generative language model *GPT4* (OpenAI 2023), which is a decoder-based model, its output is a text constructed by predicting the most likely next word, given the input sentence and the subsequently generated words. By adding to the input tweet instructions for the model, it can be conditioned

[2] This research was performed before the introduction of Twitter Blue, a paying feature allowing users to post tweets of longer length.

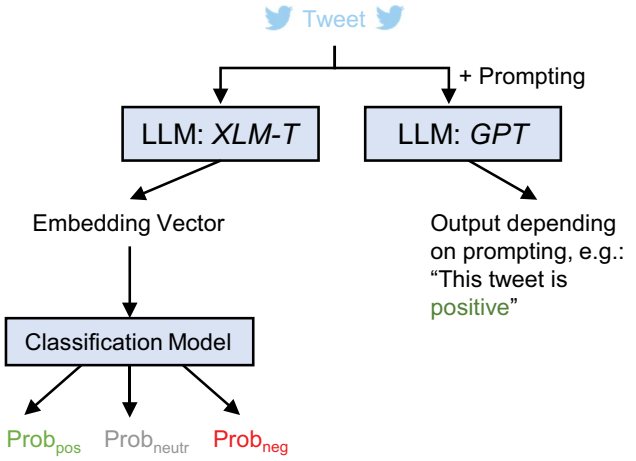


Fig. 1 Schematic view of sentiment analysis using XLM-T and GPT

to output the sentiment of the given tweet. Conditioning models with these methods is referred to as "prompting" (Liu et al. 2021). Next to their architecture, these models also differ on their training objective as well as the data used during pre-training. One of the largest differences in this is the use of Reinforcement Learning from Human Feedback for *GPT4* (Christiano et al. 2017) which is added fine-tune the model after pre-training.

For our analyses, we fist collected the tweets relevant to mobility changes in Brussels (see Sect. 3.2.1), which we then cleaned and labelled (see Sect. 3.2.2). Then, we fine-tuned *XLM-T* architecture and prompted *GPT4* for the sentiment analysis tasks (see Sect. 3.2.3).

Tweet Corpus Creation

We collected tweets through the academic research access of the Twitter API.[3] Academic research access allows users to perform "Full Archive Searches" giving the possibility to obtain any tweets since 2006. Using this access, we collected Twitter data between July 18th 2019 at 00:00 (forming of the last Brussels regional government) and December 31st 2022 at 23:59 (starting date of the analyses) on Brussels Local time (GMT + 1). Within this timeframe, five major mobility policy changes took place (in chronological order): (i) the regional Good Move plan came into force, (ii) the Brussels region became a 30km/h zone, (iii) the LTN in the city centre was announced, (iv) the LTN in the city centre was implemented, and (v) LTNs in three other municipalities were implemented.

Due to the important multilingual element of Brussels, we performed searches for tweets in three languages: French, Dutch (official languages of the region) and English. To cast the widest net possible, while limiting the collection of irrelevant tweets, we performed our search in three steps. Starting from a combination of baseline keywords relating to mobility changes, we combined these with three additional search criteria: person/official instance-based *(These are people implementing/facilitating the changes)*, location-based *(These are the places where change happens. Choosing municipalities instead of only Brussels allows for a wider net)* and keyword-based. The specificity of each search can be found in Table 1. In our search, we did not select tweets based on geolocation, as previous research showed only around 0.85% of tweets are geo-tagged (Sloan et al. 2013), limiting the pool of potential data. Additionally, since residents of Brussels are not the only ones expressing opinions about mobility changes in the city, filtering based on keywords offers a broader view. As our focus lies on the opinion of users, we excluded tweets created by accounts of media institutions, automated accounts, and known parody accounts, which were identified based on preliminary searches.

Table 1 Overview of tweet queries

Baseline Keywords: Good Move/GoodMove, Zone 30, Low Traffic Neighbourhood, Low Emission Zone		
Person/Official instances-based	**Location based**	**Keyword based**[a]
- The prime and mobility minister of the brussels regional government	- Any of the 19 brussels municipalities	- Mobility related keywords:
-Official instances related to mobility in Brussels (e.g. public transport companies)	- Neighbourhoods where low-traffic neighbourhoods have or will be implemented	Bus Metro Train Bike lane Traffic Peak Hour Car Bike Step Pedestrian
- Mayors and mobility aldermen and alderwomen of municipalities having implemented low-traffic neighbourhoods		
Language: French, Dutch and English		
Exclude: Media accounts Automated accounts Parody accounts		

[a]Previous searches included both *Good Move* and *GoodMove* as search term. This allowed some flexibility in the user's spelling of the mobility plan. As the keyword-based search is very broad, we restrict in this case the spelling to *GoodMove* to avoid false hits

Data Cleaning and Labelling

After collecting 2425 tweets, two researchers of the team independently manually labelled them over the course

[3] https://developer.twitter.com/en/use-cases/do-research/academic-research

Table 2 Obtained Cohen's Kappa (κ) interannotator agreement before (run 1) and after (run 2) debrief for the general sentiment label (κ) and the mobility change sentiment label (κ_{MC})

	Run 1 (300 Tweets)	*Run 2 (150 Tweets)*
κ	0.54	0.73
κ_{MC}	0.59	0.84

of approximately 30 h. This labelling occurred on the uncleaned tweets. For each tweet, we attributed a label from the following possibilities: Negative (0), Neutral (1), Positive (2) and Irrelevant (3). Tweets were denoted irrelevant and removed from the dataset if their subject was not related to mobility changes in Brussels. For each tweet, we labelled two sentiments: the general sentiment of the tweet and its sentiment towards the planned or implemented mobility changes.

The availability of correctly labelled data is a crucial aspect of the performance of supervised machine learning across a wide range of domains (Halevy et al. 2009; Northcutt et al. 2021). Although the traditional learning problem setting works on the assumption of noiseless and correct labels (Bootkrajang & Kaban 2011), annotation inconsistencies can occur even when labelling is performed by field experts (Sylolypavan et al. 2023). To minimize the impact of variability between the two independent annotators, we first manually labelled 300 random tweets from the dataset and computed the Cohen kappa of the labelling (Cohen 1960), which measures the agreement between annotations. A Cohen kappa value between 0.41 and 0.60 denotes a moderate agreement, while a value between 0.61 and 0.80 shows a substantial agreement between the annotators (Viera and Garrett 2005). Table 2 contains the values obtained for both the general sentiment and the sentiment towards mobility changes. The moderate values obtained for kappa indicated a difference in labelling. To remedy this, a discussion concerning the discrepancies and labelling strategies was held, focussing on the labelling strategy for sarcasm and tweets containing media titles. To evaluate this adapted strategy we again labelled 150 random tweets. The kappa value for this second round showed a more substantial agreement between the annotators for both categories. Notably, the annotators only labelled one tweet with completely opposite sentiments (i.e. positive and negative). This indicates the disagreement stemmed primarily from one annotator labelling tweets as neutral while the other assigned positive or negative labels. The remaining 2275 uncleaned tweets were then annotated independently by both researchers.

Pre-trained Language Models

After cleaning the dataset, we fine-tune *XLM-T* (Barbieri et al. 2022), a pre-trained multilingual LLM based on *XLM-R(oBERTa)* checkpoints. We use *XLM-T* due to its multilingual capabilities and because it has been further pre-trained on 198 M multilingual tweets and fine-tuned for sentiment analysis. Due to the limited amount of data we collect, with respect to the size of the original training dataset of *XLM*, using *XLM-T* is essential as it has already been trained on the intricacies of Twitter data. To analyse the sentiment of our dataset, we accessed *XLM-T* (Barbieri et al. 2022) checkpoints through the Huggingface API,[4] which we then further fine-tuned for our particular task. during 10 epochs, with a batch size of 64 and a learning rate of $5 \cdot 10^{-6}$. We also employ a polynomial learning rate scheduler, activating it after 50 warmup steps. In order to train and evaluate our model, we split the dataset into a train (75%), validation (15%) and test dataset (15%), using the later to report performances.

In an addition to fine-tuning an encoder-based LLM, we also test the capabilities of *GPT4* by OpenAI (OpenAI 2023), one of the most powerful LLM that exists at the time of writing. Instead of being fine-tuned, *GPT4* successes are often attributed to the fact that they can be conditioned for a task by either being presented examples and instructions of the task (few-shot) or by only adding instructions (zero-shot) (Kojima et al. 2023). Few-shot prompting relies on the model's capability for few-shot learning (Brown et al. 2020), where the model is fed instructions in natural language together with a number of (few) examples. However, recently the potential of zero-shot methods, where the model is only fed instructions about the task, has been demonstrated (Kojima et al. 2023), in particular in the context of reasoning tasks. For our task, we employ a zero-shot prompting method for text classification presented in Sun et al. (2023).

Results

After cleaning and labelling, we obtained a total of 1998 tweets, originating from 895 unique users. This means that, even with careful planning of data collection using the Twitter API, 16% of tweets were deemed irrelevant. Other research using twitter data over varying fields also found post-collection processing to be necessary (e.g. (Xia et al. 2021) found 56.6% of collected tweets to be irrelevant when gauging perception in context of the USA election, (Dahal et al. 2019) 7% when studying climate change related tweets,

[4] https://huggingface.co/

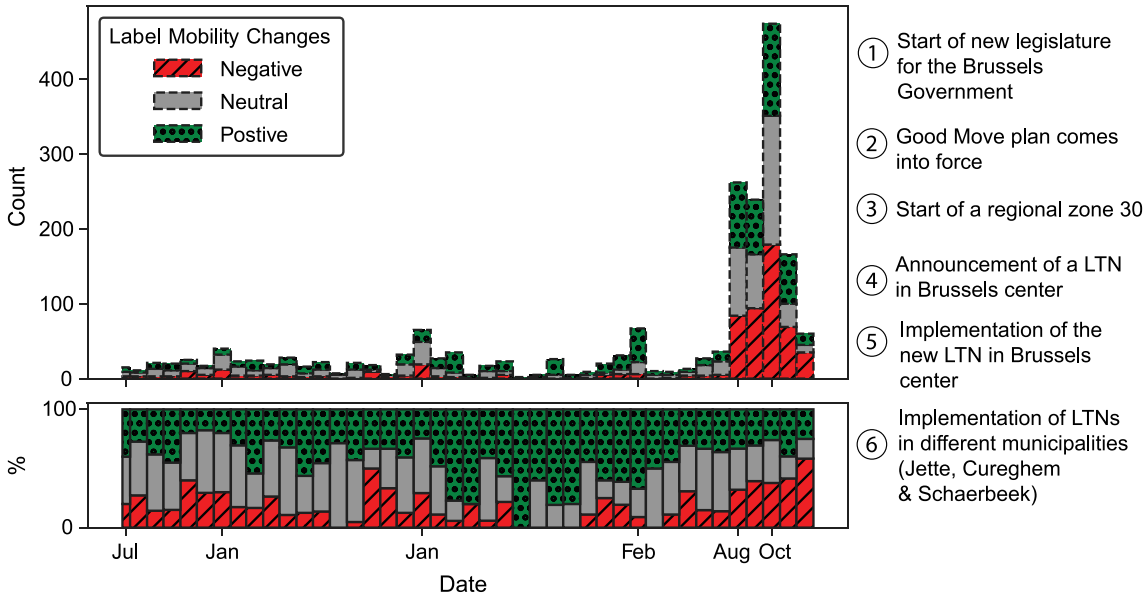


Fig. 2 Distribution of the obtained tweets by date of creation and relative proportion of the label for each time slice (months). The column for July 2019 is smaller in width to represent the data collection starting on the 19th instead of considering a full month

or 24.8% when measuring sentiment on airline services (Wan and Gao 2015). This already indicates that pipelines implementing an analysis of UGC related to specific themes require manual controls. Figure 2 shows that our dataset is heavily skewed towards tweets posted between August and October 2022, correlating with high-impact mobility changes, which generated considerable commotion (The Brussels Times 2022).

From the distribution of labels of the sentiment regarding the mobility changes, we can see that the increase in tweets posted is not due to an increase in tweets with only a negative sentiment. Another way of looking at our dataset is by analysing the evolution of the sentiment distribution over time. Figure 2 also shows that the distribution remains primarily stable, except between August and October 2022, where tweets containing negative sentiments increase percentage wise. For other notable interventions, such as the implementation of a general zone 30 km/h (January 2021) and the announcement/approval of an LTN in the centre of Brussels (October 2022), we see that tweets with negative sentiments towards mobility changes do not dominate the increase in absolute counts. By looking at the percentual distribution we can also notice that the announcement of a LTN (event number 4 on Fig. 2) generated less negative responses than the implementation of that LTN (event number 5).An essential aspect of our labelling is the difference between a tweet's sentiment and the sentiment expressed towards mobility changes in the tweet, which do not necessarily correspond. This is the case in around 30% of our dataset. A possible example of such a situation is a tweet which expresses joy when mobility changes are halted or turned back. In this case, although the tweet sentiment is positive, it represents a negative sentiment towards the (planned) mobility changes. Two concrete examples of such a tweet are displayed in Table 3 and Table(SM) 1 in the appendix contains an example for each quadrant of the matrix displayed in Fig. 3. Figure 3 shows that, for our dataset, this occurs more significantly when the tweet's sentiment is negative. In contrast, positive tweets correlate with a positive sentiment towards mobility changes more often. These discrepancies between labels form an additional difficulty when using language models for automatic labelling, as the sentiment towards the mobility changes is the most relevant for policymakers. However, correctly classifying the

Table 3 Example tweets for which the sentiment (S) and the mobility changes sentiment (S_{MC}) do not match

Tweet	S	S_{MC}
12 meetings organised for residents and local merchants in #Cureghem to co-create #GoodMove plan. All this went to waste as violence and agression fuelled by politicians made headlines and pushed the work back to the starting point… #toxicmasculinity'	0	2
(Own Translation) #BadMove The acts finally follow the decision to remove the "good move" device set up in Cureghem. The concrete blocks are being removed on this morning of Friday, September 30. The rest should follow. Congratulations to the people of Cureghem!)	2	0

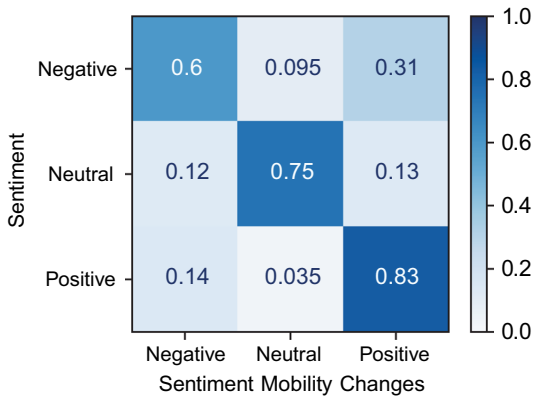


Fig. 3 Confusion matrix for the labelled tweet sentiments and sentiment expressed towards the mobility changes

Table 4 Performance scores of *XLM-T* for the three different tasks, in a zero-shot evaluation as well as after domain-specific fine-tuning

Task	Training	F1-Score	Accuracy
Sentiment	Zero-shot	42.17	0.48
	Fine-tuned	66.64	0.67
Mobility Changes (w. context)	Zero-Shot	43.47	0.39
	Fine-tuned	57.53	0.58
Mobility Changes (no context)	Zero-Shot	42.33	0.35
	Fine-tuned	56.83	0.58

tweets based on this label requires implementing an understanding of the context in the model.

General Sentiment

First, we analysed the model's performance when classifying the general sentiment of the tweets. The results of the different optimisation methods are shown in Table 4. Even though *XLM-T* is a model which has been pre-trained and fine-tuned for a sentiment analysis task, we see that the performance when no domain-specific fine-tuning has occurred is fairly low. The confusion matrices in Fig. 4 show that this poor performance is due to the model classifying tweets as neutral more often than our ground truth labelling, which occurs to a greater extent for tweets labelled as negative.

Once we trained the model on domain-specific data, we obtained an accuracy of 0.67. This indicates that, together with the zero-shot performance, although *XLM-T* has been fine-tuned for sentiment analysis on tweets, it tends to label tweets as neutral when it has not been presented the domain-specific training data.

Using the F1 score of the model, we can compare its performance with the performance of *XLM-T* on a general benchmark multilingual dataset where an average F1 score of 69.35 is obtained (Barbieri et al. 2022). Our results in Table 4 indicate that our model and training procedure performs within expectations considering the domain-specific and multilingual aspect of our task.

Mobility Change Sentiment

In a second phase, we applied *XLM-T* to classify the sentiment of the same tweets, but this time specifically towards the mobility changes, using the same training procedure as before. To correctly identify this sentiment in the tweets, context knowledge is essential. Results obtained when applying *XLM-T* in a zero-shot approach confirm this, as the accuracy is close to random guessing (see Table 4). Even after training, we can observe that the model does not obtain the same performance as on the original sentiment task. If we only consider the tweets in the test dataset whose sentiments do not match, we see that *XLM-T* reaches an accuracy of 37%, far below the accuracy on the entire dataset. The most considerable difficulty for our pretrained model when labelling is then finding the implicit context from a tweet.

Additionally, some tweets in the dataset were attributed a 'context' label during manual annotation, indicating the

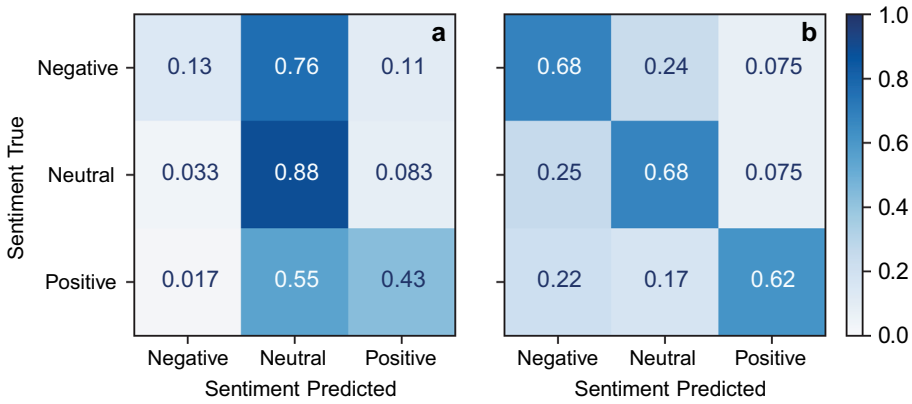


Fig. 4 Confusion matrix of the true and predicted label, for the test set, **a** when applying the model in a zero-shot way. **b** Fine-Tuning the entire XLM-T model

presence of a URL, image or other external addition, which was deemed necessary to correctly label the sentiment of the tweets towards the mobility changes. An example of this is an image showing a street containing heavy traffic, where the tweet text is (paraphrased) "Thanks GoodMove!". To assess the dependency of the model on this category of tweets, we repeat the process of fine-tuning and evaluating while removing these tweets from the dataset. The model's marginal performance increases (see Table 4) indicate that challenges with labelling are not solely attributable to tweets requiring contextual information.

Sentiment Analysis Using GPT

As *GPT* is based on a decoder architecture, it can generate any text as a response to a task, in contrast with *XLM-T*, which outputs only class labels. Using this capability, we labelled the tweets using a novel approach. Instead of classifying the tweets into three distinct categories (positive, neutral and negative) we prompt *GPT3.5* and *GPT4* to attribute each tweet a score which is a number between -1 and 1. This score reflects how negative (when closer to -1) or positive (when closer to 1) the tweet's sentiment towards the mobility plan is. An important comment is that these scores do not represent a confidence level in the sentiment, but are a value to quantify the intensity of the sentiment expressed. Adding this dimension offers a nuanced and less binary classification when performing sentiment analysis, which is more in line with how (dis)satisfaction is expressed by citizens. To compare the performance of the models we then translated these scores to labels. The classification task we perform is enhanced using Clue And Reasoning Prompting (CARP) (Sun et al. 2023), which yielded state of the art performances on text-classification benchmarks. CARP prompting enhances the ability of the model by asking it to construct an answer containing clues and a reasoning on which it then bases itself to determine the sentiment. We also explicitly mention in the prompt that the model should classify the tweets based on their opinion towards the mobility changes.

Our results (see Table 5) show that *GPT4* largely outperforms *GPT3.5* and *XLM-T* when classifying sentiments towards the mobility changes, obtaining an accuracy of 0.66. This difference in performance is even more pronounced when compared to the zero-shot performance of *XLM-T*, which obtained an accuracy of 0.39. We can also see that an accuracy of 0.66 comes close to the accuracy of *XLM-T* when classifying the general sentiment (Table 4), an easier task (as demonstrated by the accuracy of 0.58 it obtains when classifying mobility sentiment) for which it was fine-tuned. This demonstrates the potential of *GPT-4* to be used to classify implicit sentiment, a crucial aspect of UGC data related to transport and mobility changes.

Table 5 Accuracy scores for the three models XLM-T, GPT3.5-Turbo and GPT4 on the mobility sentiment classification task. Mismatched tweets are tweets whose intrinsic sentiment and sentiment towards the mobility changes do not match

Model	XLM-T (Zero-Shot)	XLM-T (Fine-Tuned)	GPT3.5-Turbo	GPT4
Accuracy	0.39	0.58	0.58	0.66
Accuracy on mismatch	0.18	0.37	0.31	0.40

Additionally, the accuracy of *GPT4* on tweets with a mismatched sentiment indicates that it might be better suited to obtain an implicit sentiment expressed in a text. This is illustrated in the response provided by *GPT4* when correctly labelling a mismatched tweet, shown in Table 6. Although the models recognizes the negative emotions expressed in the tweet, it correctly identifies the underlying positive sentiment towards the mobility changes and attributes it a positive score. This deduction capability of contextualised information is in line with the proposition that *GPT4* shows "sparks of general intelligence" (Bubeck et al. 2023).

Looking at the score distribution (Fig. 5), we observe that both *GPT3.5* and *GPT4* underuse the ranges [-0.4, -0.1] and [0.1, 0.4] to score the sentiment. Instead, neutral tweets were characterised solely by being assigned a score of 0. We note that this was purely a phenomenon originating from the models themselves, as the prompt instructed the models to use the entire range. Due to this behaviour, we translated negative scores to negative labels, zero scores to neutral labels and positive scores to positive labels. Finally, we can also see a tendency of *GPT3.5* or *GPT4* to over- or underuse certain scores, such as -0.9 for *GPT3.5* while *GPT4* avoided this score and its positive equivalent. From a machine learning perspective this raises questions about a possible underlying bias of these LLMs. This phenomenon could indicate some limitations of the model to capture/interpret certain sentiment and further investigation could prove beneficial for future tasks.

Discussion and Conclusions

Through our research, we aimed at exploring the usability of sentiment analysis through deep learning methods in transport planning, to incorporate the views of the population into decision-making.

A first important observation can be made with regards to our results. After the press coverage of the vocal negative reactions and the loud protests in the fall of 2022, more than one municipality in Brussels has paused the implementation of the regional mobility plan (BRUZZ 2022). However, when looking at the data from Twitter users, we see that the

Table 6 Tweet and response of *GPT4* after CARP prompting. Note the reasoning of *GPT4* where it classifies the tweet as positive towards the mobility changes, even though the language used initially leaned towards a negative sentiment

TWEET: (*Own Translation*)*Complete disgrace. Do you want to protest? Fine, that's your right. But stop destroying our neighbourhoods! This kind of violence must be dealt with harshly. Our security forces deserve respect ! #welovegoodmove #schaarbeek*	Label	LabelMC
	0	2
CLUES: "Complete schande" (total disgrace), "protesteren" (protest), "onze buurten kapot te maken" (destroying our neighborhoods), "geweld moet hard aangepakt worden" (violence needs to be dealt with firmly), "Onze veiligheidsdiensten verdienen respect" (our security services deserve respect), #welovegoodmove REASONING: The tweet seems to be criticizing the use of violence in protests against the GoodMove plan, calling it a total disgrace and saying it's destroying neighbourhoods. The tweeter also demands a firm response to the violence and respect for the security services. **This indicates that the tweeter is against the negative actions taken by others in opposition to the plan, but not necessarily against the plan itself, as it ends with #welovegoodmove** SENTIMENT: Positive SCORE: 0.6		

overall sentiment of our Twitter population is predominantly positive towards the implementation of the sustainable mobility plan, which provides a more nuanced perspective than the press coverage. This is contrary to our expectations, as other studies have found that social media is mainly used to express negative sentiment (Jalonen 2014). Although the sentiment of Twitter users is certainly not representative of the sentiment of the whole Brussels population, the implicit assumption that citizens are against the Good Move mobility plan because there were public outcries cannot be verified and should be nuanced. We also saw from our results that there is a correlation between the number of tweets available and the mobility interventions in the city, which shows that UGC can be an interesting and relevant complementary source of data for policy makers. For an analysis similar to ours, the initial fine-tuning of the *XLM-T* model does require time and effort, but once trained, it can then be used on multiple occasions. And although our analysis was limited in the number of tweets, we demonstrated the feasibility by comparing it to a ground truth which originates from manual labelling. However, future work can replicate this in contexts where manual labelling is not possible, i.e. with larger datasets, as we have shown that current LLMs are already quite powerful.

From our results, we see that using *GPT* offers a good alternative to provide an analysis without the need for training, since *GPT4* obtained the highest accuracy when classifying the sentiment of tweets with regard to mobility changes in a zero-shot way. This can make this type of analysis more accessible in the context of policy making, as it removes the need for experience in training models and fine-tuning. Sentiment analysis using such models like *GPT* for policy making can therefore be more easily implemented as it removes a time consuming and costly aspect of using other pre-trained models which still require fine-tuning. Apart from practical considerations, the outperformance of *GPT-4* compared to the fine-tuned model *XML-T* for detecting implicit sentiment shows that such decoder-based models are naturally better suited for this task. With well-thought prompts, decoder models also provide more information than simple classification, rendering the output more transparent, as demonstrated in Table 6. Finally we also showed that these models can be used to attribute scores to text when performing sentiment analysis, introducing a novel dimension to these kinds of analysis.

Importantly, from our results, we can say that some level of local knowledge is needed to obtain relevant content. In the tweet selection process, for example, we used the names of some politicians, as well as specific geographic locations in Brussels. If this type of analysis is to become relevant for policymaking, enough time should be spent on the inclusion criteria for the data to be used. Selection criteria can also introduce a bias into the UGC collected, as it can be skewed

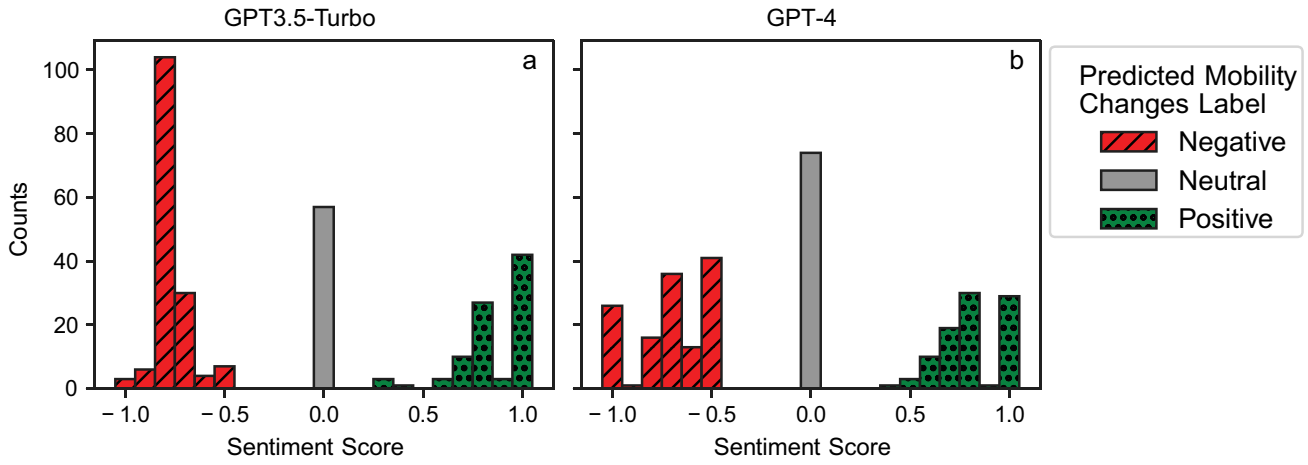


Fig. 5 Distribution of scores when classifying the sentiment of tweets with respect to the mobility changes by **a** GPT3.5-Turbo **b** GPT4. Tweets with a score of 0 were considered as neutral, between [0.1, 1] as positive and scores and between [− 1,− 0.1] as negative

towards data expressing more of a certain sentiment due to the inclusion or not of certain keywords. For policy makers, local knowledge is therefore crucial in the data selection process, to provide a holistic view on a problem.

It should also be noted that, although social media provides access to a larger dataset than could be collected through traditional sources, it does tend to exclude some users, e.g., older people (Nikolaidou and Papaioannou 2018) and people with low digital skills or no access to the internet. Additionally, the sentiments expressed on Twitter are limited to social media users and may not fully represent the broader public, since social media users are not a randomized sample of the population. It is therefore important to complement data from social media with other data sources as well, both other types of UGC, as well as data collected using traditional methods, to ensure a broad representativity for policymaking purposes.

A peculiar aspect of our dataset was the presence of two distinct sentiments in the tweets, one inherent to the vocabulary used in the tweet, the other the implicit sentiment expressed towards specific mobility changes. While a majority of the tweets collected had matching sentiments, around 30% expressed a different sentiment towards the mobility changes than one would extract from the vocabulary used. These tweets formed a difficult hurdle to overcome for all models. *GPT-4* obtained an accuracy moderately above random chance on those tweets, in contrast with its performance on the total test dataset. *XLM-T* obtained similar performances on these types of tweets, but only after pre-training the model with domain-specific data. Our results, therefore, show that detecting contextual sentiment expressed in a text is a task for which pre-trained language models still require improvement. Future work focusing on this specific type of data could therefore yield important benefits when using NLP methods in societal contexts, where there is often mismatch between inherent and implied sentiment. Combining these advancements in language models and integrating UGC data effectively, policymakers can attain a more comprehensive comprehension of public sentiment, thereby facilitating the shift in mobility towards a sustainable future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42421-024-00090-5.

Author Contributions FT: conceptualization, formal analysis, validation, writing—original draft preparation, review and editing. ST: conceptualization, validation, writing—original draft preparation (lead), review and editing. IK: validation, review and editing. VG: formal analysis, review and editing. All authors contributed to the article and approved the submitted version.

Funding Work at the VUB was supported by the Research Council (OZR) of the VUB.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on request.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes

were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Araci D (2019) FinBERT: Financial Sentiment Analysis with Pre-trained Language Models arXiv. http://arxiv.org/abs/1908.10063

Barbieri F, Ballesteros M, Saggion H (2017) Are Emojis Predictable?. arXiv. http://arxiv.org/abs/1702.07285

Barbieri F, Camacho-Collados J, Ronzano F, Espinosa Anke L, Ballesteros M, Basile V, Patti V, Saggion H (2018) SemEval 2018 Task 2: Multilingual Emoji Prediction. In: Proceedings of The 12th International Workshop on Semantic Evaluation, 24–33. https://doi.org/10.18653/v1/S18-1003

Barbieri F, Anke LE, Camacho-Collados J (2022) XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. arXiv. http://arxiv.org/abs/2104.12250

Bootkrajang J, Kaban A (2011) Multi-class Classification in the Presence of Labelling Errors. Computational Intelligence.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Amodei D (2020) Language Models are Few-Shot Learners. arXiv. http://arxiv.org/abs/2005.14165

BRUZZ. (2022). Good Move: 'Stop de democratie van de luidste roepers'. *BRUZZ*. https://www.bruzz.be/opinie/good-move-stop-de-democratie-van-de-luidste-roepers-2022-10-11

Bruzz (2020a) In kaart: Nieuwe woonerven, fietsstraten en fietspaden in Brussel. https://www.bruzz.be/mobiliteit/kaart-nieuwe-woonerven-fietsstraten-en-fietspaden-brussel-2020-05-07

Bruzz (2020b) Voor-en tegenstanders betogen in Ter Kamerenbos: 'Dit is een park, hier horen geen auto's'. https://www.bruzz.be/videoreeks/dinsdag-17-november-2020/video-voor-en-tegenstanders-betogen-ter-kamerenbos-dit-een-park

Bruxelles Mobilité. (2020a). *Plan régional de mobilité 2020–2030. Plan stratégique et opérationnel*. https://mobilite-mobiliteit.brussels/sites/default/files/2021-04/goodmove_FR_20210420.pdf

Bruxelles Mobilité (2020b) Un prestigieux prix européen pour Good Move, le nouveau plan régional de mobilité de Bruxelles. Bruxelles Mobilité. https://mobilite-mobiliteit-brussels.prezly.com/un-prestigieux-prix-europeen-pour-good-move-le-nouveau-plan-regional-de-mobilite-de-bruxelles

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee Y T, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y (2023) Sparks of Artificial General Intelligence: Early experiments with GPT-4 arXiv. http://arxiv.org/abs/2303.12712

Camacho-Collados J, Doval Y, Martínez-Cámara E, Espinosa-Anke L, Barbieri F, Schockaert S (2020) Learning Cross-lingual Embeddings from Twitter via Distant Supervision arXiv. http://arxiv.org/abs/1905.07358

Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46. https://doi.org/10.1177/001316446002000104

Collins C, Hasan S, Ukkusuri S (2013) A novel transit rider satisfaction metric: rider sentiments measured from online social media data. J Publ Trans 16(2). https://doi.org/10.5038/2375-0901.16.2.2

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised Cross-lingual Representation Learning at Scale arXiv. http://arxiv.org/abs/1911.02116

Dahal B, Kumar SAP, Li Z (2019) Topic modeling and sentiment analysis of global climate change tweets. Soc Netw Anal Min 9(1):24. https://doi.org/10.1007/s13278-019-0568-8

Derczynski L, Ritter A, Clark S, Bontcheva K (2013) Twitter Part-of-speech tagging for all: overcoming sparse and noisy data. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), 198–206.

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv. http://arxiv.org/abs/1810.04805

Endarnoto SK, Pradipta S, Nugroho AS, Purnama J (2011) Traffic condition information extraction & visualization from social media twitter for android mobile application. In: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, 1–4. https://doi.org/10.1109/ICEEI.2011.6021743

European Environment Agency. (2022). Transport and environment report 2021 [Publication]. https://www.eea.europa.eu/publications/transport-and-environment-report-2021

Evans-Cowley JS, Griffin G (2012) Microparticipation with social media for community engagement in transportation planning. Trans Res Record 2307(1):90–98. https://doi.org/10.3141/2307-10

Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1615–1625. https://doi.org/10.18653/v1/D17-1169

Gal-Tzur A, Grant-Muller SM, Kuflik T, Minkov E, Nocera S, Shoor I (2014) The potential of social media in delivering transport policy goals. Transp Policy 32:115–123. https://doi.org/10.1016/j.tranpol.2014.01.007

Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't Stop Pretraining: Adapt Language Models to Domains and Tasks arXiv. http://arxiv.org/abs/2004.10964

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12. https://doi.org/10.1109/MIS.2009.36

Haustein S, Kroesen M (2022) Shifting to more sustainable mobility styles: a latent transition approach. J Transp Geogr 103:103394. https://doi.org/10.1016/j.jtrangeo.2022.103394

Howard J, Ruder S (2018) Universal Language Model Fine-tuning for Text Classification arXiv. http://arxiv.org/abs/1801.06146

Hubert M, Corijn E, Neuwels J, Hardy M, Vermeulen S, Vaesen J (2017) From pedestrian area to urban project: Assets and challenges for the centre of Brussels (J. Corrigan, Trans.). Brussels Studies. La Revue Scientifique Pour Les Recherches Sur Bruxelles / Het Wetenschappelijk Tijdschrift Voor Onderzoek over Brussel / The Journal of Research on Brussels. https://doi.org/10.4000/brussels.1563

Hubert M (2008) L'Expo 58 et le « tout à l'automobile ». Brussels Studies. La revue scientifique pour les recherches sur Bruxelles / Het wetenschappelijk tijdschrift voor onderzoek over Brussel / The Journal of Research on Brussels. https://doi.org/10.4000/brussels.621

IBSA. (2022). Évolution annuelle | IBSA. https://ibsa.brussels/themes/population/evolution-annuelle

Jalonen H (2014) Social Media—An Arena for Venting Negative Emotions. Online Journal of Communication and Media Technologies,

4(October 2014-Special Issue), 53–70. https://doi.org/10.30935/ojcmt/5704
Jennings W, Saunders C (2019) Street demonstrations and the media agenda: an analysis of the dynamics of protest agenda setting. Comp Pol Stud 52(13–14):2283–2313. https://doi.org/10.1177/0010414019830736
Jones P (2014) The evolution of urban mobility: the interplay of academic and policy perspectives. IATSS Research 38(1):7–13. https://doi.org/10.1016/j.iatssr.2014.06.001
Khan Z, Ludlow D, Loibl W, Soomro K (2014) ICT enabled participatory urban planning and policy development: the UrbanAPI project. Trans Govern 8(2):205–229. https://doi.org/10.1108/TG-09-2013-0030
Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2023) Large Language Models are Zero-Shot Reasoners arXiv. http://arxiv.org/abs/2205.11916
Kuflik T, Minkov E, Nocera S, Grant-Muller S, Gal-Tzur A, Shoor I (2017) Automating a framework to extract and analyse transport related social media content: the potential and the challenges. Trans Res Part c 77:275–291. https://doi.org/10.1016/j.trc.2017.02.003
Lample G, Conneau A (2019) Cross-lingual Language Model Pretraining arXiv. http://arxiv.org/abs/1901.07291
Lindenau M, Böhler-Baedeker S (2014) Citizen and stakeholder involvement: a precondition for sustainable urban mobility. Trans Res Proc 4:347–360. https://doi.org/10.1016/j.trpro.2014.11.026
Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach arXiv. http://arxiv.org/abs/1907.11692
Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2021) Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing arXiv. http://arxiv.org/abs/2107.13586
Liu B (2015) Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139084789
Lock O, Pettit C (2020) Social media as passive geo-participation in transportation planning—how effective are topic modeling & sentiment analysis in comparison with citizen surveys? Geo-Spatial Inform Sci 23(4):275–292. https://doi.org/10.1080/10095020.2020.1815596
Luong TTB, Houston D (2015) Public opinions of light rail service in Los Angeles, an analysis using Twitter data.
Macharis C, Tori S, de Séjournet A, Keseru I, Vanhaverbeke L (2021) Can the COVID-19 crisis be a catalyst for transition to sustainable urban mobility? Assessment of the Medium- and Longer-Term Impact of the COVID-19 Crisis on Mobility in Brussels. Front Sustain 2. https://doi.org/10.3389/frsus.2021.725689
Martin-Domingo L, Martín JC, Mandsberg G (2019) Social media as a resource for sentiment analysis of Airport Service Quality (ASQ). J Air Transp Manag 78:106–115. https://doi.org/10.1016/j.jairtraman.2019.01.004
Maynard D, Funk A (2012) Automatic detection of political opinions in tweets. In: R. García-Castro, D. Fensel, G. Antoniou (Eds.), The Semantic Web: ESWC 2011 Workshops (pp. 88–99). Springer. https://doi.org/10.1007/978-3-642-25953-1_8
Milmo D (2007) Protest greets congestion charge's westward push. The Guardian. https://www.theguardian.com/environment/2007/feb/20/localgovernment.greaterlondonauthority
Naseem U, Razzak I, Khushi M, Eklund PW, Kim J (2021) COVIDSenti: a large-scale benchmark twitter data set for COVID-19 sentiment analysis. IEEE Trans Comput Soc Syst 8(4):1003–1015. https://doi.org/10.1109/TCSS.2021.3051189
Nikolaidou A, Papaioannou P (2018) Utilizing social media in transport planning and public transit quality: survey of literature. J Trans Eng Part a 144(4):04018007. https://doi.org/10.1061/JTEPBS.0000128
Northcutt CG, Athalye A, Mueller J (2021) Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks arXiv. http://arxiv.org/abs/2103.14749
O'Sullivan F (2017) Barcelona's Car-Taming Superblock Plan Faces a Backlash. Bloomberg. https://www.bloomberg.com/news/articles/2017-01-20/barcelona-s-superblocks-expand-but-face-protests
OpenAI (2023) GPT-4 Technical Report arXiv. http://arxiv.org/abs/2303.08774
Pagolu VS, Reddy KN, Panda G, Majhi B (2016) Sentiment analysis of Twitter data for predicting stock market movements. In: 2016 international conference on signal processing, communication, power and embedded system (SCOPES), 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659
Parry IWH, Walls M, Harrington W (2007) Automobile externalities and policies. J Econ Literature 45(2):373–399. https://doi.org/10.1257/jel.45.2.373
Pucci P, Vecchio G (2019) Big Data: hidden challenges for a fair mobility planning. In: P. Pucci & G. Vecchio (Eds.), Enabling Mobilities: Planning Tools for People and Their Mobilities (pp. 43–58). Springer International Publishing. https://doi.org/10.1007/978-3-030-19581-6_4
Quan C, Ren F (2016) Visualizing emotions from chinese blogs by textual emotion analysis and recognition techniques. Int J Inf Technol Decis Mak 15(01):215–234. https://doi.org/10.1142/S0219622014500710
Rietzler A, Stabinger S, Opitz P, Engl S (2019) Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification arXiv. http://arxiv.org/abs/1908.11860
Rotmans J, Loorbach D, Kemp R (2012) Complexity and transition management. Routledge, In Complexity and Planning
Semanjski I, Bellens R, Gautama S, Witlox F (2016) Integrating big data into a sustainable mobility policy 2.0 planning support system. Sustainability, 8(11), Article 11. https://doi.org/10.3390/su8111142
Serna A, Gerrikagoitia JK, Bernabé U, Ruiz T (2017) Sustainability analysis on urban mobility based on social media content. Trans Res Proc 24:1–8. https://doi.org/10.1016/j.trpro.2017.05.059
Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, Rana O (2013) Knowing the tweeters: deriving sociologically relevant demographics from twitter. Sociol Res Online 18(3):74–84. https://doi.org/10.5153/sro.3001
Small KA, Verhoef ET (2007) The economics of urban transportation. Routledge
Stopher PR, Greaves SP (2007) Household travel surveys: Where are we going? Trans Res Part a 41(5):367–381. https://doi.org/10.1016/j.tra.2006.09.005
Sun X, Li X, Li J, Wu F, Guo S, Zhang T, Wang G (2023) Text Classification via Large Language Models. arXiv. http://arxiv.org/abs/2305.08377
Sylolypavan A, Sleeman D, Wu H, Sim M (2023) The impact of inconsistent human annotations on AI driven clinical decision making. Npj Digital Medicine 6(1):26. https://doi.org/10.1038/s41746-023-00773-3
The Brussels Times. (2022). Good Move? Mobility plans provoke fierce protest as drivers fear losing freedom. https://www.brusselstimes.com/317057/mobility-plans-provoke-fierce-protest-as-motorists-fear-loss-of-freedom
Tuarob S, Tucker CS (2015) Quantifying product favorability and extracting notable product features using large scale social media data. J Comput Inform Sci Eng 15(3). https://doi.org/10.1115/1.4029562

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017). Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363

Wan Y, Gao Q (2015) An ensemble sentiment classification system of twitter data for airline services analysis. IEEE Int Conf Data Min Workshop (ICDMW) 2015:1318–1325. https://doi.org/10.1109/ICDMW.2015.7

We Are Social & Meltwater (2023) Digital 2023 October Global Statshot Report. We Are Social Global. https://wearesocial.com/blog/2023/10/digital-2023-october-global-statshot-report/

Widener MJ, Li W (2014) Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Appl Geogr 54:189–197. https://doi.org/10.1016/j.apgeog.2014.07.017

Wyrwoll C (2014) *User-Generated Content. In Social Media Fundamentals, Models, and Ranking of User-Generated Content* (pp 11–45). Springer Vieweg https://doi.org/10.1007/978-3-658-06984-1_2

Xia E, Yue H, Liu H (2021) Tweet sentiment analysis of the 2020 U.S. presidential election. Companion Proc Web Conf 2021:367–371. https://doi.org/10.1145/3442442.3452322

Xu H, Liu B, Shu L, Yu PS (2019) BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis arXiv. http://arxiv.org/abs/1904.02232

Zannat KE, Choudhury CF (2019) Emerging big data sources for public transport planning: a systematic review on current state of art and future research directions. J Indian Inst Sci 99(4):601–619. https://doi.org/10.1007/s41745-019-00125-9

Zhang H, Zhang Y, Zhan L-M, Chen J, Shi G, Wu X-M, Lam AYS (2021) Effectiveness of Pre-training for few-shot intent classification. Findings of the Association for Computational Linguistics: EMNLP 2021, 1114–1120. https://doi.org/10.18653/v1/2021.findings-emnlp.96

Zhang H, Liang H, Zhang Y, Zhan L-M, Wu X-M, Lu X, Lam A (2022) Fine-tuning Pre-trained Language Models for Few-shot Intent Detection: Supervised Pre-training and Isotropization. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 532–542. https://doi.org/10.18653/v1/2022.naacl-main.39

Zipori E, Cohen MJ (2015) Anticipating post-automobility: design policies for fostering urban mobility transitions. Int J Urban Sustain Dev 7(2):147–165. https://doi.org/10.1080/19463138.2014.991737

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Floriano Tori[1] · Sara Tori[2] · Imre Keseru[2] · Vincent Ginis[1,3,4]

✉ Floriano Tori
floriano.tori@vub.be

[1] Data Analytics Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[2] Mobilise Research Group, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[3] Applied Physics Group, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[4] School of Engineering and Applied Sciences, Harvard University, Cambridge, USA