Annotations for Exploring Food Tweets From Multiple Aspects

Matīss Rikters[λ], **Edison Marrese-Taylor**[λ], Rinalds Vīksna[ι]
[λ] National Institute of Advanced Industrial Science and Technology, [ι] University of Latvia

Introduction

We build upon the **Latvian Twitter Eater Corpus** (LTEC), which is focused on the narrow domain of tweets related to food, drinks, eating and drinking and has been collected for more than 12 years and reaching almost 3 million tweets with the basic information as well as extended automatically and manually annotated metadata. In this paper we supplement the LTEC with manually annotated subsets of evaluation data for **machine translation** (MT), **named entity recognition** (NER), timeline-balanced **sentiment analysis**, and **text-image relation** classification. We experiment with each of the data sets using baseline models and highlight future challenges for various modelling approaches. The data sets are published on **https://github.com/Usprogis/Latvian-Twitter-Eater-Corpus**.

Named Entity Recognition

The test set was annotated in the CoNLL-2003 format with four entity classes (PER, LOC, ORG, MISC). Cohen's Kappa between two main annotators for the task was 0.92. The test set contains 188 MISC entities, 99 LOC entities, 55 PER entities, and 68 ORG entities. We trained 2 **BERT-based NER models** using data from other sources and found that pre-training on the Latvian food tweets performs better overall.

Entity	mBERT	mBERT+tweets
LOC	88.44	**94.00**
MISC	**85.25**	80.75
ORG	75.00	**83.82**
PER	85.22	**89.66**
Overall	84.41	**85.71**

Machine Translation

Manual translation into English was performed by a translator and afterwards reviewed by a post-editor as quality assurance. We evaluated open-source models and publicly available translation services using the dataset and compared automatic evaluation scores according to BLEU.

	BLEU	ChrF
Google Translate	43.09	65.75
Tilde MT	**48.28**	**68.21**
Opus MT	41.73	62.38
mBART	33.36	54.45

Text-image Results

We experimented with **LLaVA models** prompting them using the original tweet texts in Latvian or automatic translations into English.

Pr. 1	Given the following text, extracted from a tweet in Latvian: *Brokastis ēd! Nu ēd, ēd!!!* Is the image adding to the text meaning? Reply "Yes" or "No".
Pr. 2	Given the following text, extracted from a tweet in Latvian: *Brokastis ēd! Nu ēd, ēd!!!* Is the text represented in the image? Reply "Yes" or "No".

Prediction accuracy was **20.69%** when evaluated on the original texts, and improved slightly to **27.83%** on English translations.

	Gold	Predicted	Correctly
		Latvian	
🖼️✅ 📄✅	392	240	113
🖼️✅ 📄❎	72	442	**35**
🖼️❎ 📄✅	296	31	**8**
🖼️❎ 📄❎	52	99	12
		English	
🖼️✅ 📄✅	392	336	**172**
🖼️✅ 📄❎	72	379	**35**
🖼️❎ 📄✅	296	20	6
🖼️❎ 📄❎	52	77	**13**

🖼️ – image adds to the text meaning, 📄 – text is represented in the image, ✅ – true, ❎ – false.

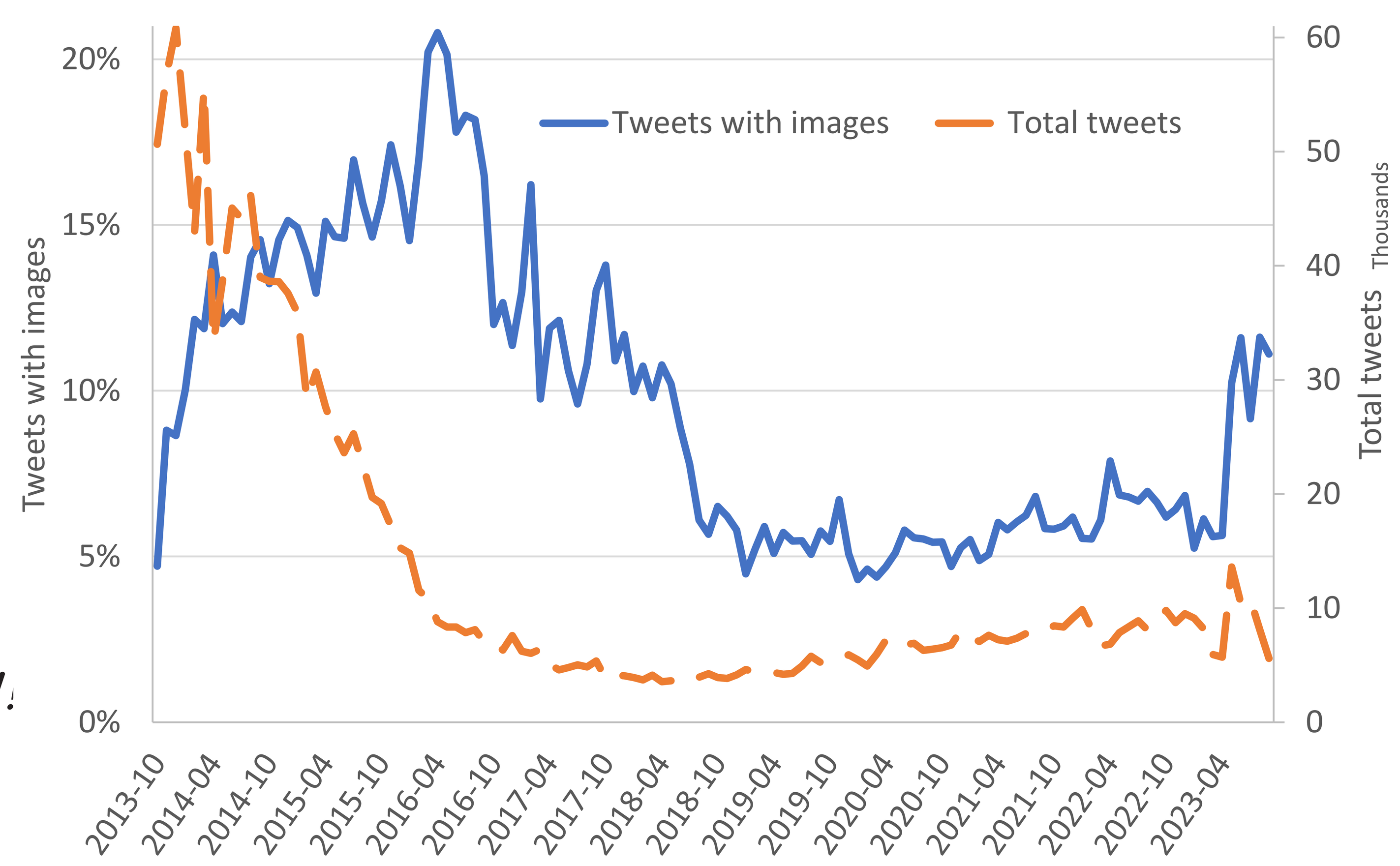
Annotation Environments

For MT and NER, we manually translated and annotated the same **test set of 744 tweets**, which already had manually assigned sentiment classes. However, this was too poor in terms of attached images. For the text-image relation task we sampled **800 recent tweets with image attachments**. For the timeline-balanced sentiment analysis test set, we sampled 50 random tweets from each year **between 2011 and 2020, totalling 500**. Annotators for NER and the text-image relation tasks were native Latvian speakers in their 20s-30s with at least a master's degree and work experience in the field of natural language processing or linguistics. For sentiment analysis, in 20s-50s with at least a bachelor's degree, and the translator and post-editor were professional linguists with a master's degree in their respective field in their 30s-40s.

Text-image Relation

LV	*Brokastis ēd! Nu ēd, ēd!!!*
EN	*Eat breakfast! Well, eat, eat!!!*

The popularity of posting images along with food-related tweets has shifted over the years **between 5–20% of total monthly tweets**. After annotating 800 image-tweet pairs we found that the majority do indeed textually describe what is represented in the image, and ∼1/2 of the cases the image also adds to the meaning of the text. Only ∼9% add to the meaning without describing the contents in the text, and a mere 6% do neither.



Sentiment Analysis

12 evaluators individually judged sentiment for 500 tweets. We used the majority vote of the evaluators as the final annotation in cases with disagreement, and considered two classifications as correct in 21 cases where the majority opinion was split equal. Overall agreement of the evaluators was 70.48% with a free marginal kappa of 0.56 (0.40 to 0.75 considered intermediate to good agreement). We trained a **BERT-based classifier** using 20k training tweets from other sources and evaluated on the two test sets, reaching 74.06% accuracy on the original 744 tweet set and 86.40% on the timeline-balanced set. The accuracy of the average evaluator compared to the majority was 80.25%.