Annotations for Exploring Food Tweets From Multiple Aspects

Matīss Rikters[λ], Edison Marrese-Taylor[λ], Rinalds Vīksna[ι]

[λ] Artificial Intelligence Research Center,
National Institute of Advanced Industrial Science and Technology
{firstname.lastname}@aist.go.jp
[ι] University of Latvia
{rinaldsviksna}@gmail.com

Abstract

This research builds upon the Latvian Twitter Eater Corpus (LTEC), which is focused on the narrow domain of tweets related to food, drinks, eating and drinking. LTEC 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, named entity recognition, 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.

Keywords: Social Media Analysis, Corpus Creation, Multimodality

1. Introduction

Despite the recently induced chaos due to the company leadership change, Twitter (now renamed to X[1]) has long been and still remains one of the most influential social networks not only around political or technological topics, but also for everyday lifestyle content and regular people posting about their daily lives. For years Twitter was also one of the only remaining useful social media platforms to the research community by providing real-time access to new posts with additional elevated access for academic research purposes at no cost[2]. This, however, also became deprecated in fall of 2023[3] with no reasonable alternative so far, aside from the $5000.00 USD/month Pro tier. The Latvian Twitter Eater Corpus (LTEC (Sproģis and Rikters, 2020)) is a collection of tweets gathered by following the appearance of 363 keywords related to food and eating inflected in various valid word forms in the Latvian language. Data collecting was started in 2011 and has reached over 2.5 million tweets generated by more than 170,000 users. Each tweet in the dataset is represented by standard data fields like tweet date, text and id, author screen name, as well as additional metadata fields where available, such as location details, attached media information, a list of food and drink products mentioned in the tweet text in original surface forms and nominative singular forms (with respective English translations), and a separate subset of tweets with manually annotated sentiment classes - positive, neutral and negative.

In this paper we describe supplementing the LTEC with several new manually annotated and task-specific evaluation sets for text-image relation classification, machine translation, named entity recognition and an additional set for sentiment analysis. We also provide results from baseline experiments using each of these evaluation sets

2. Related Work

The LTEC has been analysed from several aspects of sensory experiences related to the food products mentioned in the tweet content by Kāle et al. (2021), as well as comparing the change of sentiment in food tweets over time. Most recently (Kāle and Rikters, 2023) linked the timeline of food tweet sentiment changes with weather observation data in the area for further insights. We believe that our additions to the dataset can further spark analytical research of food-related tweet data from other aspects, such as the images related to tweets.

Vempala and Preoţiuc-Pietro (2019) explore the relations between Tweet text and attached images and categorise them in four different groups: a) the image adds to the text meaning and the text is represented in the image; b) the image adds to the text meaning and the text is not represented in the image; c) the image does not add to the text meaning and the text is represented in the image; and d) the image does not add to the text meaning and the text is not represented in the image. They also further analyse user demographic traits linked to each of the four image tweeting types.

---

[1] From Twitter to X: Elon Musk Begins Erasing an Iconic Internet Brand - https://www.nytimes.com/2023/07/24/technology/twitter-x-elon-musk.html

[2] Guide to the future of the Twitter API - https://developer.twitter.com/en/products/twitter-api/early-access/guide

[3] https://developer.twitter.com/en/updates/changelog

3. Manually Annotated Datasets

In this section, we discuss the newly created manually annotated parts of LTEC. For the machine translation (MT) and named entity recognition (NER) tasks, we manually translated and annotated the same test set of 744 tweets, which already had manually assigned sentiment classes. This test set, however, was far too poor in terms of attached images, so for the text-image relation task we sampled 800 recent tweets with image attachments. Finally, for the extended and more balanced sentiment analysis test set, we sampled 50 random tweets from each year between 2011 and 2020, totalling 500.

The number of annotators was task-specific, having 12 annotators for sentiment analysis, a total of 3 for NER and image-text relation, and 2 for translation. With some of them overlapping between tasks, there were a total of 16 unique annotators. A more detailed annotator profile is provided in the appendix along with annotation environment descriptions and annotation instructions.

3.1. Text-image Relation

While not available from the very beginning of Twitter, the option to embed images directly into a tweet was added in 2011 (CNN, 2011) after Twit-Pic had been the main external source of attached images for several years. The popularity of posting images along with food-related tweets has shifted over the years between 5 and 20 per cent of total monthly tweets, as can be seen in Figure 2. We use the same annotation schema as Vempala and Preoțiuc-Pietro (2019) by grouping image-text relations into whether the image adds to the text meaning and whether the text is represented in the image. We manually annotate a test set of 800 image – tweet pairs. In our annotations we found that the majority do indeed textually describe what is represented in the image, and about half of the cases the image also adds to the meaning of the text. Only around 9% add to the meaning without describing the contents in the text, and a mere 6% do neither.

For example, given the image shown in Figure 1 and its accompanying text from the tweet (and its English translation), it is fairly clear that the text does not at all describe what is shown in the image, since we can see a cat sitting on a chair in the image and it is not mentioned anywhere in the text. However, the image most certainly adds to the meaning of the text – upon seeing the angry look of the cat it becomes clear that the text represents its personified thoughts.

3.2. Machine Translation

Manual translation of the 744 tweet test set from the original Latvian language into English was per-

| Text | *Brokastis ēd! Nu ēd, ēd!!!* |
| Translation | *Eat breakfast! Well, eat, eat!!!* |

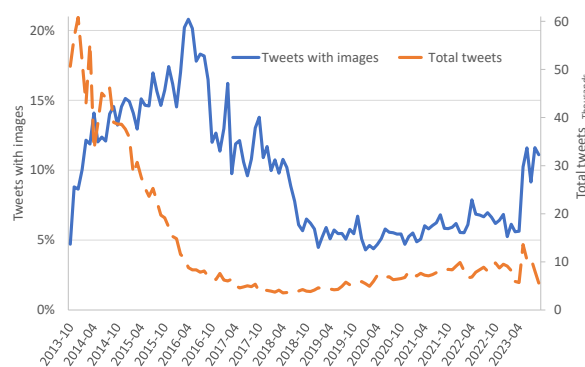Figure 1: The attached image to the tweet text mentioned in Table 1.

Figure 2: The total count of monthly tweets contrasted to the percentage of monthly tweets containing images.

formed by a single translator. After the translation the sentences were reviewed by a professional post-editor as an additional quality assurance measure. Such texts tend to be challenging to comprehend and convey by modern models due to the use of specific terminology and abbreviations, which at times are used to maintain posts within certain character count limitations of social networks.

3.3. Named Entity Recognition

The 744 tweet test set was annotated in the CoNLL-2003 format (Tjong Kim Sang and De Meulder, 2003) with four entity classes for persons, locations, organisations and miscellaneous (PER, LOC, ORG, MISC), mostly following MUC-7 guidelines (Chinchor, 1998). Annotations

were performed with a reduced entity set similar to the Balanced State-of-the-Art Multilayer Corpus for natural language understanding[4] (Gruzitis et al., 2018) by converting the nine entity tag set to a four entity tag set. We kept the PERSON and ORGANIZATION tags, joined the LOCATION and GPE tags into a single LOCATION tag, and joined the remaining tags into a single MISC category. The manual annotation process was carried out by 2 annotators using the INCEpTION (Klie et al., 2018) tool, with a third annotator resolving conflicts. Cohen's Kappa (as reported by the INCEpTION) between two main annotators for the NER task was 0.92. The test set contains 188 MISC entities, 99 LOC entities, 55 PER entities, and 68 ORG entities. We also introduced a new FOOD tag to be used for marking food and drink related entities.

3.4. Sentiment Analysis

The original sentiment analysis test set in LTEC was sampled from tweets of only one year while the whole corpus spans over a decade. While translations and named entities are somewhat more independent from the specific time period, sentiment can be more time-sensitive. For a more trustworthy evaluation, we decided to randomly select 50 tweets per year from each year between 2011 and 2020 into a more balanced additional 500 tweet evaluation set. Twelve human evaluators were asked to individually judge the sentiment for each of the 500 tweets. We used the majority vote of the human evaluators as the final annotation in cases where they disagreed on a particular evaluation and considered two classifications as correct in the 21 cases where the majority opinion was split in half (for example, 6 positive and 6 neutral). The overall agreement of the evaluators was 70.48% with a free marginal kappa (Randolph, 2005) of 0.56 (values from 0.40 to 0.75 are considered intermediate to good agreement). This shows that the tweet texts are not always trivial enough to be unequivocally classified into just one of the three sentiment classes.

4. Experiments

4.1. Text-image Relation

Many modern multi-modal large language models have recently gained the ability to answer questions about a given image in plain text chat form. In the text-image relation experiments we feed our data to the LLaVA (Liu et al., 2023) model[5] and prompt it to answer a question about the type of text-image relation using the two prompts outlined

---

	Prompt 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".
	Prompt 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".

Table 1: Examples of the prompts used for querying the LLaVA model.

	Gold	Predicted	Coorectly
		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**

Table 2: Automatic prediction results by LLaVA. 🖼 represents the image adding to the text meaning, 📄 – the text being represented in the image, and ✅ and ❌ – true or false respectively.

in Table 1. We experimented with providing the model with the original tweet texts written in Latvian, as well as automatically translated versions of tweet texts into English using Tilde MT[6], since the LLaVA model is using Vicuna (Zheng et al., 2023) as its language model part, which is inherently not multilingual. Prediction accuracy was only 20.69% when evaluated on the original texts, and improved slightly to 27.83% on English translations. The results in Table 2 show more detailed results, which outline how the model tends to mostly answer the first question positively at around 84-88% of the time, while it should be around 57%. Meanwhile, the model seems to be answering the second question overly negatively at around 56-67%, where the answer should be closer to 15%, since in the vast majority of annotated tweets the content of the text was indeed represented in the image.

4.2. Machine Translation

We evaluated several state of the art MT models using the dataset and compared automatic evaluation according to BLEU (Papineni et al., 2002)

	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

Table 3: Machine translation experiment results.

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**

Table 4: Results (F$_1$ scores) of applied NER.

and ChrF (Popović, 2015) scores using sacre-BLEU (Post, 2018). We compare the translations from publicly available translation services such as Google Translate[7] and Tilde MT[8], and open source pre-trained models like Opus MT (Tiedemann and Thottingal, 2020) and mBART (Tang et al., 2021). The results summarised in Table 3 indicate that such social media data is not particularly challenging to translate. While mBART does noticeably fall behind the others three, BLEU scores in the range of 40 to 50 usually indicate fairly usable translations. Upon manual inspection there was one very noticeable frequent error made by the pre-trained models which was absent in the output from translation services and is directly related to typical modern social media texts – partial or full absence of emoji and emoticons.

4.3. Named Entity Recognition

The task of named entity recognition (NER) has been proven more challenging for languages with less widely available data, and especially morphologically rich languages such as Latvian. Recognising named entities (NEs) in texts from social media adds an extra layer to the challenge, as such texts are not necessarily the most grammatically correct, and even tend to use specific terminology or abbreviations to maintain the content of the post within the limits of the specific social media network (for Twitter formally 140, later changed to 280 symbols).

We trained 2 NER models using (Grūzītis et al., 2019) dataset[9]: a fine-tuned multilingual BERT (mBERT (Devlin et al., 2019)) and fine-tuned mBERT additionally pre-trained on tweets and added emoticons vocabulary[10] (Thakkar and Pinnis, 2020) using the Flair library (Schweter and Akbik, 2020).

The results in Table 4 show that a language model additionally pre-trained on Latvian tweets performs better when applied to in-domain data in a downstream task. The overall F$_1$ score

is lower than SOTA (88.1) (Ulčar and Robnik-Šikonja, 2020) on the Fullstack (Grūzītis et al., 2019) dataset, however, the LitLat score is obtained performing the NER on only 3 classes: person, location, organisation.

4.4. Sentiment Analysis

With using the manually annotated 5,420 tweet training set to fine-tune a pre-trained multilingual BERT (Devlin et al., 2019) model for the sentiment analysis task along with ∼20,000 sentiment-annotated Latvian tweets from other sources[11], the 744 tweet test set provided in LTEC can reach an accuracy of around 74%. The accuracy of the model according to the majority of human evaluators on our timeline-balanced 500 tweet set was even higher, reaching 86.40%. Since our annotation was performed by 12 annotators who did not always agree, we also compared each of their annotations against the majority vote. The accuracy of the average human evaluator compared to the majority was 80.25%, which does not necessarily indicate human parity of the model, but rather that the correct choice is not always easy to decide.

5. Conclusion

In this paper, we introduced an extended version of evaluation sets for the Latvian Twitter Eater Corpus by supplementing the current sentiment analysis test set with named entity labels and full translations into English. We also manually labelled two new evaluation sets for understanding relations that attached images have with the posted text, as well as a more timeline-balanced sentiment analysis test set.

By experimenting with evaluating pre-trained models using our new test sets we find that some tasks like named entity recognition and sentiment analysis can already handle such data fairly well. However, other tasks like machine translation and especially image-text relation classification still leaves much room for improvement.

---

[7]Google Translate October 2023 version `https://translate.google.com`

[8]Tilde MT October 2023 version - `https://translate.tilde.com/`

[9]`https://github.com/LUMII-AILab/FullStack`

[10]`https://huggingface.co/FFZG-cleopatra/bert-emoji-latvian-twitter`

[11]Other Latvian twitter sentiment corpora - `https://github.com/Usprogis/Latvian-Twitter-Eater-Corpus/tree/master/sub-corpora/sentiment-analysis#other-latvian-twitter-sentiment-corpora`

We plan to publish the newly annotated data and merge them into the main repository of LTEC[12], keeping everything under the current MIT license.

Limitations

In this work, we only considered highly domain-specific data written in a relatively less-spoken language, and possibly containing a noticeable degree of internet slang. Therefore, any results obtained from analysing such data need to be interpreted with these considerations in mind and cannot be generalised to the broader scope of the Latvian language. Also, since hyper-parameter tuning on training large models is computationally very costly, we opt for choosing mostly default parameters and base versions of open-source models for our experiments.

Ethics Statement

Our work fully complies with the ACL Code of Ethics[13]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. All human annotators were fairly compensated for their efforts.

Acknowledgments

6. Bibliographical References

Nancy A. Chinchor. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Doug Gross CNN. 2011. Twitter launches own photo-sharing tool. `http://edition.cnn.com/2011/TECH/social.media/08/10/twitter.photos/index.html`. Accessed: 2023-10-19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Normunds Grūzītis, Lauma Pretkalniņa, Baiba Saulīte, Laura Rituma, Gunta Nešpore-Bērzkalne, Pēteris Paikens, Ilze Auziņa, Artūrs Znotiņš, Kristīne Levāne-Petrova, and Roberts Darģis. 2019. Full stack of latvian language resources for NLU. CLARIN-LV digital library at IMCS, University of Latvia.

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018. Creation of a balanced state-of-the-art multilayer corpus for NLU. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Maija Kāle and Matīss Rikters. 2023. What food do we tweet about on a rainy day? In *The 29th Annual Conference of the Association for Natural Language Processing (NLP2023)*.

Maija Kāle, Jurģis Šķilters, and Matīss Rikters. 2021. Tracing multisensory food experiences on twitter. *International Journal of Food Design*, 6(Gastrophysics):181–212.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

---

[12]Latvian Twitter Eater Corpus - `https://github.com/Usprogis/Latvian-Twitter-Eater-Corpus/`

[13]`https://www.aclweb.org/portal/content/acl-code-ethics`

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the Joensuu Learning and Instruction Symposium*, volume 2005.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition.

Uga Sproģis and Matīss Rikters. 2020. What Can We Learn From Almost a Decade of Food Tweets. In *In Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020)*, volume 328, pages 191 – 198, Kaunas, Lithuania. IOS Press.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Gaurish Thakkar and Mārcis Pinnis. 2020. Pre-training and fine-tuning strategies for sentiment analysis of latvian tweets. *Human Language Technologies – The Baltic Perspective (Baltic HLT 2020)*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Lit-Lat BERT. CLARIN-LT digital library in the Republic of Lithuania.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

7. Language Resource References