Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200622

What Can We Learn from Almost a Decade of Food Tweets

Uga SPROĢIS [a,1], Matīss RIKTERS [b]
[a] *Faculty of Computing, University of Latvia, Latvia*
[b] *The University of Tokyo, Japan*

Abstract. We present the Latvian Twitter Eater Corpus - a set of tweets in the narrow domain related to food, drinks, eating and drinking. The corpus has been collected over time-span of over 8 years and includes over 2 million tweets entailed with additional useful data. We also separate two sub-corpora of question and answer tweets and sentiment annotated tweets. We analyse the contents of the corpus and demonstrate use-cases for the sub-corpora by training domain-specific question-answering and sentiment-analysis models using the data from the corpus.

Keywords. Annotated corpora, social networks, food data, Latvian

1. Introduction

Even though the usage and popularity of Twitter have stopped rapidly growing and even dropped in recent years[2], it still has a considerable amount of loyal users who keep on sharing everything from worldwide events to random personal details with their followers. We decided to focus on one of the random personal details that people share, specifically, anything to do with food consumption and related topics.

Several corpora of Latvian tweets exist in prior work, but none of them are domain-specific and have been collected over an extensive period of time. Milajevs [1] collected and analysed 1.4 million tweets geo-located in Riga, Latvia from April 2017 to July 2018 and 60 thousand tweets [2] from November 2016 to March 2017. Pinnis [3] collected and analysed 3.8 million tweets of Latvian politicians, companies, media, and users who interacted from August 2016 to July 2018 There are also several data sets of general sentiment-annotated tweets [4], [5], [3][3] amounting to 14,781 tweets in total.

In this paper, we describe the Twitter eater corpus (TEC) and analyse its contents. We also provide two sub-corpora: one consisting of question and answer tweets and one with sentiment-annotated tweets. More details can be found in Section 2. In Sections 3.1 and 3.2, we describe question answering and sentiment analysis experiments using our corpus. Finally, we conclude the paper in Section 4.

[1] Corresponding Author: Uga Sproģis; E-mail: ugasprogis12@inbox.lv.
[2] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users
[3] https://github.com/nicemanis/LV-twitter-sentiment-corpus

2. The Twitter Eater Corpus

The corpus consists of tweets that have been collected from October 2011 [6] until April 2020. They are tracked using 363 keywords, which are various inflections of Latvian words associated with eating, tasting, breakfast, lunch, dinner, etc. The main keywords are shown in Table 1: the words in bold are mostly verbs that describe eating - these were inflected to all usable forms and included in the full keyword list. The rest of the keywords are a set of the top 60 food-related words that were the most popular in the first month of collecting the tweets.

Figure 1 illustrates the contents of a single tweet from the TEC in JSON notation. Each tweet consists of primary fields - *"tweet_id"*, *"tweet_text"*, *"tweet_author"* and *"created_at"*, which will always be present, and optional fields, which depend on the tweet text and metadata. We separate three groups of optional fields: 1) "*media_url*" and *"expanded_url"*, which contain information about the media files from the tweet; 2) *"location_name"*, *"location_lng"*, *"location_lat"* and *"location_country"*, which specify where the tweet was created; and 3) *"food_surface_form"*, *"food_nominative_form"*, *"food_group"* and *"food_english_translation"*, which contain semicolon-separated lists of foods or drinks that appear in the tweet.

At the beginning of the project, approximately 15,000 food and drink words from collected tweets were manually annotated with their respective nominative forms, English translations and food groups according to the food guide pyramid [7]. The food groups are: bread, cereal, rice, pasta (6); vegetables (5); fruit, berries (4); milk products (3); meat, eggs, fish (2); fats, oils, sweets (1). There are two additional groups for drinks: alcoholic drinks (7) and non-alcoholic drinks (8).

The corpus is available on Github[4], in accordance with the content redistribution section of the Twitter Developer Agreement and Policy[5]. The public release includes tweet IDs along with data fields created within the scope of this project (starting with *"location_lng"* in Figure 1). The complete version is available upon individual request for research purposes. The repository also includes data processing scripts and details on how to reproduce our experiments.

Table 1. List of main keywords used to collect the corpus

taste	**lunch**	beet	potato	mandarin	sweet
eat	**feast**	bun	cabbage	sauce	mushroom
breakfast	**drink**	carrot	candy	pancake	onion
dine	**treat**	chips	sour cream	dumpling	chocolate
dinner	**nom**	vegetable	cream soup	gingerbread	tea
bite	**appetite**	meat	cake	rice	tomato
meal	orange	Hesburger	drink	salad	grape
food	apple	coffee	McDonald's	ice cream	strawberry

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

[5]https://developer.twitter.com/en/developer-terms/agreement-and-policy

```
{
  "tweet_id": 1213025400273735680,
  "tweet_text": "Gulašzupa #receptesĪsumā gulašzupa ir gana
     vienkārša liellopu gaļas bāzēta zupa https://t.co/
     OnqDwotQr0 https://t.co/Z2tAodyj9M",
  "tweet_author": "receptes_eu",
  "created_at": "2020-01-03 11:12:54",
  "media_url": "http://pbs.twimg.com/media/ENWIKb8WsAAiLKE.
     jpg",
  "expanded_url": "https://twitter.com/receptes_eu/status/12
     13025400273735680/photo/1",
  "location_name": "Ogresgals",
  "location_lng": "24.7377",
  "location_lat": "56.8079",
  "location_country": "Latvia",
  "food_surface_form": "Gulašzupa;liellopu;gaļas;zupa;",
  "food_nominative_form": "gulašs;liellops;gaļa;zupa;",
  "food_group": "2;2;2;6;",
  "food_english_translation": "Goulash;Cattle;Meat;Soup;",
}
```


Figure 1. An example of a tweet from the TEC with all available metadata

2.1. Content Overview

The corpus contains 2,275,787 tweets, of which 155,057 contain media information, 165,335 contain location information and 1,297,159 tweets mention foods or drinks. Table 2 shows the 10 most popular foods and drinks from the TEC. Looking from a Latvian consumer perspective[6], it is very typical that Latvians mostly drink water, tea, juice, beer and eat meat, vegetables and fruits. Interesting, however, is the high popularity of sweets such as chocolate, cakes, ice cream and Coca-Cola.

Table 2. List of foods and drinks which are the most popular overall

Food	Count	Drink	Count
Chocolate	117,235	Tea	163,338
Ice cream	86,109	Coffee	120,040
Meat	85,574	Juice	18,179
Potatoes	70,135	Water	15,692
Salads	61,616	Beer	14,845
Cake	52,267	Cocktails	8,207
Soup	46,545	Coca-cola	5,016
Pancakes	40,203	Alcohol	4,766
Sauce	40,201	Champagne	3,673
Apple	36,571	Vodka	2,802

[6]https://enciklopedija.lv/skirklis/4980-nacion%C4%811%C4%81-virtuve-Latvij%C4%81

Figure 2 shows the yearly count of collected tweets along with the potential trend (since for 2011 and 2020, only a part has been collected) and the general popularity of Twitter and Instagram (a competing social network) for Latvia from Google Trends [7]. There was a stable income of food tweets up until 2015, however, it seems that the following decrease correlates with the overall drop in the popularity of Twitter in Latvia, which seems to be directly opposite to the popularity of Instagram in Latvia according to Google Trends.

In Figure 3, we have visualised four of the largest tweet trends over the past years from the Latvian speaking twitter users. The most recent one just a month ago - panic buying of buckwheat due to the CoViD19 pandemic of 2020, followed by the doubling of butter prices in 2017, Latvian sprat import ban to Russia in 2015, and, finally, the horsemeat scandal in 2013. If we look closer at the 2823 tweets about meat in week 9 of 2013, we can see multiple inflexions of the word "horse" along with words like "scandal" and "investigation" among the most common words.

Figure 4 shows a selection of seasonal trends averaged from data between 2012 and 2019. Most trends have one peak zone indicating parts of the year when they are more popular. Examples of this are gingerbread and tangerines in December, and strawberries and ice cream in the summer. We were expecting to see chocolate peak high on Valentine's day, but while it does peak, the difference is not as high.

2.2. *Question - Answer Sub-corpus*

We noticed that there are a number of tweets in our corpus that express questions. To highlight one of the uses of the corpus, we selected a subset of tweets which include

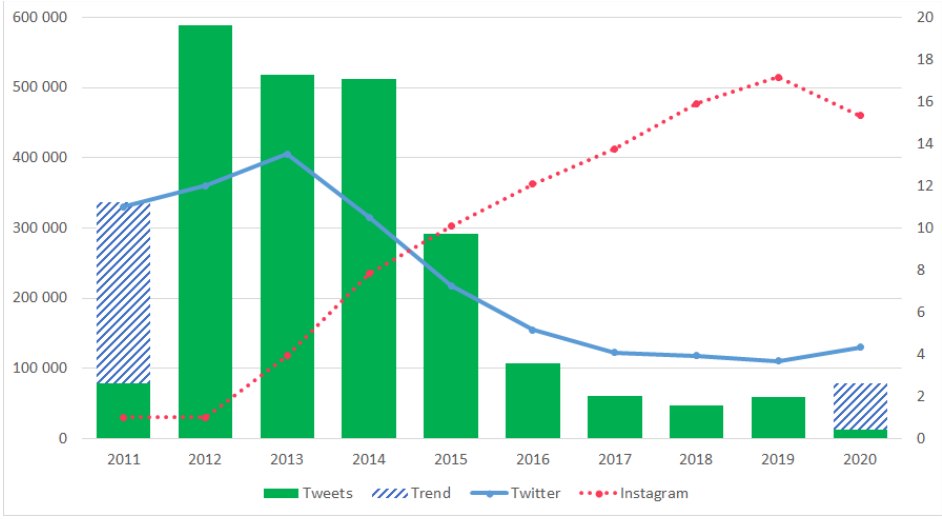

Figure 2. Collected tweet count by year

[7] https://trends.google.com/trends/explore?hl=en-US&tz=-540&date=2011-10-06+2020-03-14&geo=LV&q=%2Fm%2F0fjd36,%2Fm%2F0289n8t,%2Fm%2F02y1vz,%2Fm%2F0glpjll&sni=3

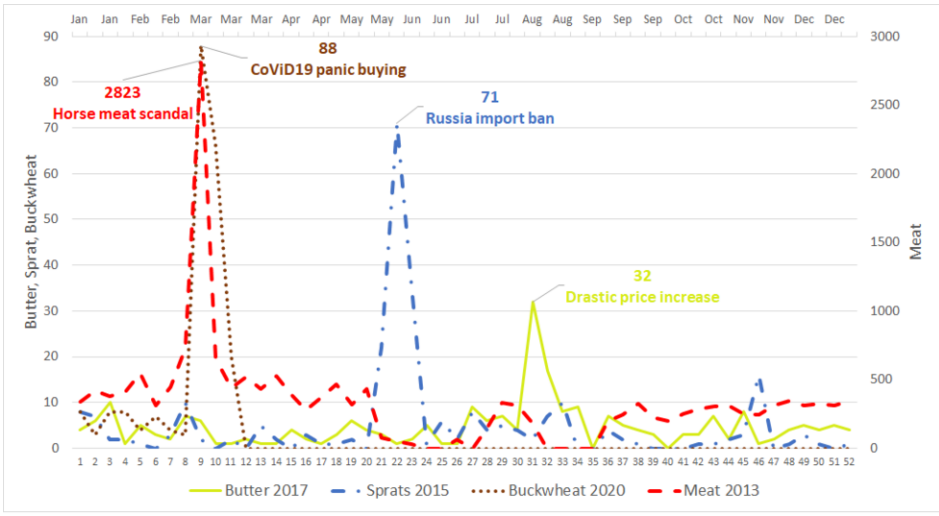

Figure 3. Four of the large trends noticeable in the TEC

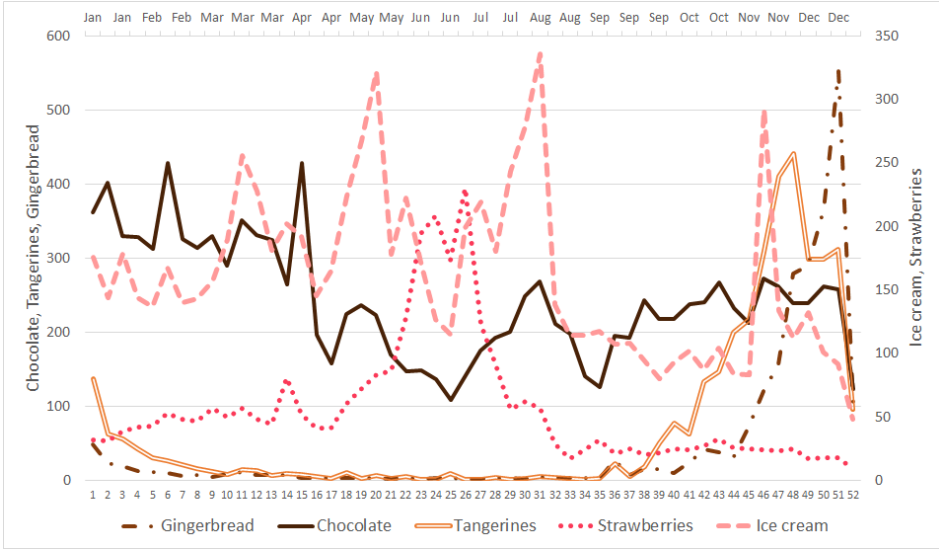

Figure 4. Five of the yearly seasonal trends noticeable in the TEC

at least one of typical Latvian question words[8] or phrases along with a question mark. This resulted in 215,233 question tweets. To gather answers for them, we scraped Twitter's web version[9], which resulted in 19,871 tweets with at least one reply. Since there were many tweets with multiple answers, we eventually wound up with 42,744 question-answer pairs. We randomly selected subsets of 1,000 and 500 question-answer pairs to use as the development set and evaluation set respectively.

[8]http://valoda.ailab.lv/latval/vidusskolai/SINTAKSE/sint3jaut.htm
[9]https://github.com/luodaoyi/TwEater

2.3. Sentiment Annotated Sub-corpus

We manually annotated 5,420 tweets, marking them as positive, neutral or negative. This gave us 1,631 positive, 2,507 neutral and 1,282 negative tweets. We further split these into a test set of 250 tweets from each class and a training set.

3. Experiments

3.1. Question Answering

Typical question answering systems are trained using paragraphs of text, questions about the paragraphs and answers to those questions [8]. Since we only had question-answer, we chose to train an encoder-decoder model similar to machine translation using questions and answers as source and target languages, respectively. We used Sockeye [9] to train transformer architecture models with the base parameters until they reached convergence on development data.

Our initial experiments using only TEC data showed rather poorly generated answers due to the lack of general-domain training data. To mitigate this, we used the same approach to select question-answer tweets from the Latvian Tweet Corpus [3]. This gave us 546,982 additional question-answer pairs to add to our training data.

3.1.1. Data Pre-processing

We performed tokenisation or truecasing using scripts from the Moses Toolkit [10]. We used Sentencepiece [11] to create a shared subword vocabulary of 8,000 tokens. We replaced all Twitter-specific *@user* mentions with *@USR* and URLs with *@URL*, as these usually do not contain relevant linguistic data for the model to learn. We also replaced multiple consecutive *@USR* or *@URL* tags with a single one and removed them completely if they were either at the start of the tweet or in the end.

3.1.2. Results

Figure 5 shows several examples of generated answers to the evaluation data questions. There were many hypothesis answers that were not even close to the reference ones but still made some sense in relation to the questions, such as the first two. There were also just as many or even more answers that made no sense at all like the last one.

We performed a small-scale human evaluation on the results by asking 5 annotators to evaluate a random 10 % of the evaluation set by marking generated answers as either OK or not good (NG). The evaluators marked 46.40 % of answers as OK. The evaluators had an overall agreement of 66.27 % (Free-marginal kappa [12] - 0.33), which indicates moderate agreement.

3.2. Sentiment Analysis

For the sentiment analysis experiment, we used 3,116 of our sentiment-annotated tweets. 250 random tweets from each sentiment class were separated as an evaluation corpus and the remaining 2,368 were used for training.

Question: Bet es jau tevis dēļ visu daru. Cikos jāiet ēst torti?
Hypothesis: Es jau paēdu.: D
Reference: uzzvanīšu kad varēs nākten ciemen.

Question: Man šķūnī nesen piedzima ežuļi. Tāpēc SOS jautājums - ko ēd mazi eži?
Hypothesis: Jā, tas ir ļoti labs jautājums.:)
Reference: tikai nebaro ar krūti, viņi nepanes pienu!

Question: kāpostu tīteņi vai cepelīni?
Hypothesis: Jā, svaigēdājus.
Reference: kāpostu tīteņi

Figure 5. Example questions with generated hypothesis answers and true reference answers

For sentiment analysis, we performed similar data pre-processing steps as for question answering, except for splitting words in sentence pieces. We also experimented with stemming[10] and lemmatizing [13] words.

Table 3 shows the results of our sentiment analysis experiments. We compared a Python implementation of the Naive Bayes classifier from NLTK[14] against Pinnis [3] implementation of the Perceptron classifier. We also experimented with several combinations of training data sets - TE (our Twitter Eater dataset), MP [3], RV [5], PE [4], NI[11]. We found that the highest classification accuracy - 61.23 % - is achieved by using all but NI data sets for training and only stemming all words.

Table 3. Accuracy of our sentiment analysis experiment results on scale of 0 to 100

Training Data	TE	MP	MP.PE	TE.MP	All	TE.MP.RV.PE
Naive Bayes	53.21	43.32	45.72	56.55	59.63	58.02
Perceptron	53.07	52.67	53.47	57.87	57.33	58.27
			Stemmed			
Naive Bayes	53.74	46.39	50.67	58.16	60.56	**61.23**
Perceptron	56.67	53.73	54.13	60.00	56.93	57.73
			Lemmas			
Naive Bayes	53.88	45.45	49.60	56.42	58.42	59.63
Perceptron	54.41	51.07	53.07	57.35	56.95	56.95
			Stemmed Lemmas			
Naive Bayes	54.41	45.99	49.33	57.62	59.63	59.63
Perceptron	53.34	51.47	52.67	58.29	56.68	57.09

[10] https://github.com/rihardsk/LatvianStemmer
[11] https://github.com/nicemanis/LV-twitter-sentiment-corpus

4. Conclusion

In this paper, we described the creation of a fairly large narrow-domain corpus of Twitter posts related to the topic of eating. We gave some insights in overall observations gained from the corpus contents and various trends that we noticed from the data. We believe that the data would be useful in many linguistic, sociological, behavioural and other research areas.

We experimented with creating a food-related question answering system using one subset of our data and a sentiment analysis system using another subset to highlight potential use-cases of our corpus. While the results did not break new ground, we hope that they inspire related future research.

Acknowledgements

We would like to thank Mārcis Pinnis for sharing his collected tweet dataset with us as well as running experiments with his model using our data.

References

[1] Milajevs D. Language use in a multilingual tweet corpus. In: Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018. vol. 307. IOS Press; 2018. p. 88.

[2] Milajevs D. Toward a Comparable Corpus of Latvian, Russian and English Tweets. In: Proceedings of the 10th Workshop on Building and Using Comparable Corpora. Vancouver, Canada: Association for Computational Linguistics; 2017. p. 26–30.

[3] Pinnis M. Latvian Tweet Corpus and Investigation of Sentiment Analysis for Latvian. In: Proceedings of the 8th Baltic Conference on Human Language Technologies (Baltic HLT 2018); 2018. p. 112–119.

[4] Peisenieks J, Skadiņš R. Uses of Machine Translation in the Sentiment Analysis of Tweets. IOS Press. 2014.

[5] Vīksna R. Sentiment Analysis in Latvian Tweets [Master's Thesis]. Rīgas Tehniskā universitāte; 2018.

[6] Rikters M. Universālas metodes Twitter datu analīzei [Bachelor's Thesis]. Latvijas Universitāte; 2012.

[7] Duston D. Food guide pyramid is built on a base of grains. Daily News. 1992 Apr:8–8.

[8] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics; 2016. p. 2383–2392.

[9] Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, et al. Sockeye: A Toolkit for Neural Machine Translation. ArXiv e-prints. 2017 dec.

[10] Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al.. Moses: open source toolkit for statistical machine translation. Association for Computational Linguistics; 2007.

[11] Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2018. p. 66–71.

[12] Randolph JJ. Free-Marginal Multirater Kappa (multirater κfree): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In: Presented at the Joensuu Learning and Instruction Symposium. vol. 2005; 2005. .

[13] Paikens P. Lexicon-based morphological analysis of Latvian language. In: Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT 2007); 2007. .

[14] Bird S, Loper E, Klein E. Natural language processing with python O'reilly media Inc; 2009.