Gaussian Processes for One-class and Binary Classification of Crisis-related Tweets

Jens Kersten*
German Aerospace Center – Jena, Germany†
jens.kersten@dlr.de

Jan Bongard
jan.bongard@dlr.de

Friederike Klan
friederike.klan@dlr.de

ABSTRACT

Overload reduction is essential to exploit Twitter text data for crisis management. Often used pre-trained machine learning models require training data for both, crisis-related and off-topic content. However, this task can also be formulated as a one-class classification problem in which labeled off-topic samples are not required. Gaussian processes (GPs) have great potential in both, binary and one-class settings and are therefore investigated in this work. Deep kernel learning combines the representative power of text embeddings with the Bayesian formalism of GPs. Motivated by this, we investigate the potential of deep kernel models for the task of classifying crisis-related tweet texts with special emphasis on cross-event applications. Compared to standard binary neural networks, first experiments with one-class GP models reveal a great potential for realistic scenarios, offering a fast and flexible approach for interactive model training without requiring off-topic training samples and comprehensive expert knowledge (only two model parameters involved).

Keywords

Gaussian Process, One-class Classification, Twitter, Overload Reduction, Crisis Informatics

INTRODUCTION

The benefits, potentials and limitations of social media data for crisis management have extensively been investigated, for example with respect to media usage patterns (Reuter and Kaufhold 2018), incident databases (Wiegmann et al. 2021), public information and warning (Zhang et al. 2019), practical challenges (Stieglitz et al. 2018), and collaborative emergency management (Fathi et al. 2020). Identifying the subset of data that is related to the specific question at hand usually is the first step in social media-based applications. However, this challenging task of information overload reduction (Hiltz and Plotnick 2013; Rao et al. 2017; Kaufhold et al. 2020) is known to be one of the main barriers to practically exploit social media data in emergency management (Plotnick and Hiltz 2016; Stieglitz et al. 2018).

The most widely employed methods for classifying disaster-related social media text messages are based on supervised (Burel and Alani 2018; Kersten et al. 2019) and semi-supervised machine learning (Kaufhold et al. 2020; Kruspe et al. 2021). In (Wiegmann et al. 2020a), type-agnostic neural network models trained with data covering 9 common disaster event types and tested with data from unseen events yielded an average F_1-score of 0.88 and an average false positive rate (FPR) of 4.7 % when applied to 5 million random off-topic tweets. Even though this indicates a good performance for a wide range of applications, drawbacks of such pre-trained models are the lack of adaptability as well as the potentially low and unknown generalization capability in case of new events and event types. To mitigate these issues, models trained from scratch utilizing *ad hoc* labeled data (Kaufhold et al. 2020;

*corresponding author
†www.dlr.de/dw/en

Snyder et al. 2020), domain adaptation approaches (Mazloom et al. 2019) and few-shot models (Kruspe et al. 2019) might be possible alternatives. User interaction and thematic model adjustment functionalities in turn require expert knowledge. Furthermore, the usually considered binary classification problem requires representative samples of unrelated tweets (the class we are actually not interested in) to learn model parameters. At the cost of lower expected F_1-scores, recently proposed few-shot models (Kruspe et al. 2019) only require few examples to detect semantically similar tweets in unseen data, and also offer one-class classification capabilities.

A further promising approach for both, classification and regression problems, is given with Gaussian processes (GPs) (Rasmussen and Williams 2005) providing a Bayesian non-parametric, computationally tractable machine learning framework. The following properties of GPs motivate the research presented in this paper: (1) GPs can easily be combined with pre-trained deep learning models (for example sentence embeddings); (2) GPs are known to be a suitable choice in case of few training samples (Deborah et al. 2017); (3) GPs can not only solve binary, but also one-class classification (OCC) problems (Kemmler et al. 2013); (4) GPs provide probabilistic predictions. These properties directly address demands in practical emergency management, where interactivity and flexibility is desired and model uncertainty information can support prediction assessment.

A further potential benefit of OCC models is related to significantly imbalanced class occurrences, especially in the case of Twitter streams that are not filtered by keywords. While balanced sample sets for training and testing classifiers help to understand model performance in general, this setting does not represent the real ratio of related and unrelated samples. Additionally, the off-topic class basically covers every possible unrelated content, which makes sampling difficult. These challenges are circumvented in OCC. Especially in case of unexpected events, discussions and developments hidden in a Twitter data stream, new OCC GP models may be trained from scratch to detect similar microblog messages solely based on a small number of positive examples.

In the field of Twitter text analysis, GPs were only occasionally investigated, for example for rumor stance classification (Lukasik et al. 2015; Lukasik et al. 2019), sentiment analysis (Deborah et al. 2017) and text regression in general (Beck 2017). Taking advantage of both, the excellent representative power of deep neural networks (DNNs) and the probabilistic framework of GPs, their combination, i.e., GPs on top of DNNs and trained end-to-end, was investigated for tasks like image classification (Bradshaw et al. 2017) and image quality assessment (Camps et al. 2018).

In line with these works, we investigate the combination of DNN-based feature representations (sentence embeddings) and GPs for the task of identifying crisis-related tweets in binary and OCC settings. To gain first insights, we examine the cross-event performance of general purpose GP models to identify any crisis-related tweet regardless of the event type. Furthermore, the question of how GPs can be used for OCC in context of crisis-related situational monitoring is investigated. The addressed tasks are therefore the classification of crisis-related tweets based on (1) binary classification models that require positive and negative training samples, and (2) OCC models that only require training samples for the class of crisis-related tweets.

RELATED WORK

Compared to often investigated neural networks (Kruspe et al. 2021) or non-probabilistic support vector machines (SVMs) (C. Burges and C. J. Burges 1998), GPs gained less attraction in NLP for social media. In (Preoţiuc-Pietro and Cohn 2013), periodic distributions of hashtag frequencies on Twitter over time are modeled in order to forecast expected hashtag volumes based on past data. Multi-task GPs are used in (Cohn and Specia 2013) to model annotator bias and for learning from outputs of multiple annotators while accounting for annotator-specific behavior. Rumor stance classification aims at classifying social media user reactions related to potential rumors. In (Lukasik et al. 2019), multi-task GPs are investigated for this task in a one-versus-all fashion.

In (Deborah et al. 2017), a bag-of-words feature representation and multi-kernel GPs are investigated for the task of Twitter sentiment analysis. Multi-kernel learning automatically identifies the most suitable subset of kernels (or combinations of these) instead of using a fixed kernel function. Genetically evolved Gaussian kernels for sentiment analysis are investigated in (Roman et al. 2019). GPs for text regression tasks in general, and for emotion analysis and machine translation quality estimation, are thoroughly addressed in (Beck 2017). Special emphasis is put on deriving kernels for text data, for example, based on hard string matching and soft embedding-based matching.

In deep kernel learning (DKL) (Wilson et al. 2016; Ober et al. 2021) the excellent representative power of DNNs and the Bayesian formalism of GPs are fused to hybrid models that can be trained end-to-end. A DNN is used to obtain a low-dimensional latent feature representation fed into a GP. As an example, experiments conducted in (Bradshaw et al. 2017) revealed, that DKL can be advantageous in transfer testing and for adversarial robustness. In analogy to DNNs, deep Gaussian processes (DGPs) (Damianou and Lawrence 2013; Jayashree and Srijith 2020) are multiple stacked GPs that are intended to learn rich representational functions along with uncertainty estimates.

To the best of our knowledge, GP models were not yet investigated for the tasks of classifying crisis-related Twitter microblogs. Related work in OCC problems focuses on other application domains, for example novelty or anomaly detection in images (Kemmler et al. 2013) or image quality assessment (Camps et al. 2018), but indicates great potential for various applications.

GP CLASSIFICATION MODELS

In this section, the concept of GPs is outlined. Furthermore, the binary and OCC models proposed and investigated for classifying crisis-related tweets are described.

Gaussian Processes

GPs provide a Bayesian non-parametric machine learning framework which unites sophisticated and consistent theory with computational tractability (Rasmussen and Williams 2005). In supervised learning, we seek to learn a model f that maps input vectors $\boldsymbol{x}$ to target values $y = f(\boldsymbol{x})$. Instead of parameterizing a fixed f explicitly, we can assume that the function is drawn from a specific probability distribution. The (latent) function f is distributed according to a GP, if and only if any finite subset of function values $f(\boldsymbol{x})$ has a joint Gaussian distribution. Hence, even though the number of possible function values is infinite, inference based on a finite subset of observations will provide the same result as if the infinite set of all unobserved points would have been taken into account.

Given a finite set of observed data samples and a GP prior that represents expected function behavior (e.g. smoothness), the desired model outputs for new observations are represented as a posterior distribution over functions f and can be obtained via Bayes' rule. For both, regression and classification problems, the inference result for a new data sample is given with the mean (and variance) of the aforementioned posterior over functions at this point. A GP is completely specified by its mean function $m(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ pairwise evaluated at observations $\boldsymbol{x}$ and $\boldsymbol{x}'$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), \kappa(\boldsymbol{x}, \boldsymbol{x}')). \tag{1}$$

Instead of using linear functions to model the relationship between $\boldsymbol{x}$ and y, the input $\boldsymbol{x}$ could be mapped into a higher N-dimensional feature space using a nonlinear function (e.g. polynomials). As long as the projections are fixed functions, the model is still linear and therefore analytically tractable (Rasmussen and Williams 2005). Given the fact that an explicit mapping can be circumvented by choosing a covariance function that is defined in terms of inner products in input space (i.e., the covariance between the outputs is defined as a function of the inputs), learning equals to find a suitable covariance function.

GPs for Regression and Classification

Regression aims to map each input vector $\boldsymbol{x}$ to a value $y \in [0, 1]$. Under the assumption of a zero-mean function $m(\boldsymbol{x})$, the moments of the predictive GP distribution are defined by[1]

$$\begin{aligned} \mu_* &= \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\ \sigma_*^2 &= k_{**} - \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_*, \end{aligned} \tag{2}$$

where $*$ marks inputs and inference results for new observations $\boldsymbol{x}_*$, $\boldsymbol{I}$ is the identity matrix and σ_n^2 is the prior variance for all n training observations $\boldsymbol{X}$. Furthermore, the abbreviations $\boldsymbol{K} = \kappa(\boldsymbol{X}, \boldsymbol{X})$, $\boldsymbol{k}_* = \kappa(\boldsymbol{X}, \boldsymbol{x}_*)$ and $k_{**} = \kappa(\boldsymbol{x}_*, \boldsymbol{x}_*)$ denote the covariance function applied on the training set $\boldsymbol{X}$ and the unseen test data $\boldsymbol{x}_*$.

A common choice for the covariance function is the squared exponential kernel of the form

$$\kappa(\boldsymbol{x}, \boldsymbol{x}_*) = \sigma^2 \exp\left(-\frac{\|x - x_*\|^2}{2l^2}\right), \tag{3}$$

where the lengthscale l controls the smoothness and σ^2 the variance of $f(x)$. Given a set of training data $[\boldsymbol{X}, \boldsymbol{y}]$, these hyperparameters can be learned by maximizing the marginal likelihood $p\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta} = \left(l, \sigma^2\right)\right)$, i.e., the probability of the data given the model, with respect to these parameters.

[1] More details on the derivation of equations can be found in (Rasmussen and Williams 2005).

As described for example in (Kemmler et al. 2013), GPs can also be applied to the special task of **one-class classification**. Due to the constant training labels $\boldsymbol{y} = \mathbf{1}$, the mean in (2) simplifies to

$$\mu_* = \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I} \right)^{-1} \mathbf{1}. \tag{4}$$

A decision whether the model prefers the positive class or not can simply be derived by thresholding the moments, or related heuristics, like $H = \mu_* \sigma_*^{-1}$ (Kapoor et al. 2010). However, compared to regression, where the model parameters can be learned with gradient-based methods, the constant training labels would cause convergence to a linear regression function that always predicts values $y \sim 1$. As a consequence, hyperparameter tuning is only possible by taking into account application domain-specific conditions and assumptions.

The goal of **binary classification** is to predict the label $y \in \{0, 1\}$ given a feature vector $\boldsymbol{x}$. In contrast to available closed-form solutions for the regression case, the discrete output space in binary classification leads to non-Gaussian posteriors. A common approach to derive a tractable solution in this case is to approximate the posterior with a normal distribution.

DATASET

The disaster tweet corpus 2020 (Wiegmann et al. 2020b) is a large compilation of various publicly available sets, comprising 48 disasters from the ten most common man-made and natural disaster types. For each tweet, a binary label indicating the relatedness to a specific disaster event is provided. The original data sets are imbalanced with a usually much higher number of related samples. Further, since all tweets were acquired by keyword search, the original unrelated tweets do not represent the full distribution of this class (compared to an unfiltered set). To address these two issues, balanced label ratios were obtained by simply replacing all initially given unrelated tweets (from the original data sets) by appropriate amounts of unrelated tweets randomly sampled from a set of 5 million *tranquil* tweets. Since the tranquil set solely contains off-topic tweets without the restriction to specific keywords that have to be contained, it can also be used to evaluate the false positive rate of filtering models. An overview of covered event types, number of events and labeled tweets is provided in Table 1.

Table 1. Event types, corresponding number of events and labeled tweets of the disaster tweet corpus.

Event Type	# Events	# Labels
Biological	2	6,106
Flood	9	14,210
Wildfire	3	4,820
Hurricane	9	48,922
Earthquake	11	27,034
Tornado	2	11,204
Transportation	4	3,850
Societal	2	11,206
Industrial	4	10,166
Other	2	2,852
Tranquil	-	5 Million

PROPOSED MODELS

The large-scale evaluation framework proposed in (Wiegmann et al. 2020a) serves as our methodological baseline and experimental setup, since the same task with special emphasis on cross-disaster experiments is addressed. Within this frame, we investigate one dedicated GP model for each task, binary and one-class classification of crisis-related tweets, with specific focus on cross-event generalization capabilities.

Baseline Model

The input layer of our baseline model for binary classification is a pre-trained universal sentence encoder (USE) (Cer et al. 2018) converting each input tweet text to a 512-dimensional, real-valued feature vector[2]. These vectors are then fed into a 3-layer feed-forward neural network (256, 128 and 2 nodes, respectively) with ReLUs as activation function, and Max-Entropy as optimization criterion. It has to be noted that several other pre-trained embeddings

[2] https://tfhub.dev/google/universal-sentence-encoder-large/5

would be a suitable choice here, especially models tailored to tweet texts, like BERTweet (Nguyen et al. 2020) or to crisis-related tweets, like CrisisBERT (Liu et al. 2020). However, for a better comparability to the baseline model, we utilize the USE embedding throughout this article. Experiments with other embeddings are planned in future works.

Binary GP Model

In the binary setting, the USE embedding vectors are fed into a sequence of two fully connected layers with 128 and 16 nodes and ReLu activations. A reduced number of 16 latent dimensions (compared to 128 nodes in the baseline DNN) is intended to keep the GP-related computational loads low. The resulting latent feature vectors are then passed to a single Gaussian process layer with a squared exponential kernel and two output dimensions. According to the discrete target values $y \in \{0, 1\}$, a likelihood layer with a Bernoulli distribution is used as final layer. End-to-end training with the Adam optimizer is conducted based on tensorflow (Abadi et al. 2015) and the GPflux library (Dutordoir et al. 2021).

One-class GP Model

The potential use case for a one-class GP model is that there are no representative training samples available, for example in case of new upcoming events or event types, new target regions or new specific target (sub-) topics of interest. By providing a set of initial representative samples of the desired microblog texts, this model seeks to find similar texts within the data stream. Since gradient-based training in this one-class case is not possible, intermediate fully connected layers are not introduced here. The USE embedding vectors are directly fed into a squared exponential kernel GP-layer (van der Wilk et al. 2020) in combination with a Gaussian likelihood layer yielding scalar outputs $y \in [0, 1]$. An input tweet text is classified as related, if $y > 0.5$. This model involves only two GP parameters, lengthscale and variance, which cannot be estimated based on training data. Their sensitivity is therefore investigated in the experimental section.

EXPERIMENTAL RESULTS AND DISCUSSION

The following research questions are addressed in two corresponding experiments: (1) Are binary GP models suited for classifying crisis-related tweets and for Twitter stream filtering? (2) How well do one-class GP models perform for these tasks and how to choose appropriate model parameters? In order to ensure comparability to the baseline model, the same splits of training and test events with 10-fold cross-validation (CV) are used. Furthermore, the tranquil set of 5 million unrelated tweets is used to evaluate the filtering capabilities in terms of false positive rates.

In each of the two 10-fold CV experiments, a general-purpose GP model is trained with a randomly sampled tweet set covering all involved disaster types. Each model is then evaluated with respect to all event types individually with test tweets from unseen events. In order to investigate the impact of the training set size, nine different training sample sizes in range $N = \{50, \ldots, 3,000\}$ are evaluated.

Binary GP Models

In each CV run, the binary GP model parameters are trained with N training samples for 50 epochs. The corresponding average F_1-scores and false positive rates are summarized in Table 2.

In case of $N = 3,000$ training samples, the baseline DNN yields slightly better F_1-scores for wildfire, transport, industrial, and societal, as well as a maximum F_1 gain of 0.044 for biological. This indicates a slightly better DNN generalization capability when training data from few events (between 2 and 4) are available. In contrast, GP F_1-scores turn out to be significantly better for event types covering more than 4 events (e.g. a gain of 0.06 for hurricanes). This indicates a better GP generalization capability compared to DNNs if training data from various events is available. Average GP F_1 values for all event types is also higher ($\Delta F_1 = 0.04$), which might be influenced by the fact that event types with better GP F_1-scores are also represented by significantly more labeled tweets.

In case of fewer training samples ($N \geq 250$), a maximum F_1 decrease of 0.064 for biological and around 0.03 for all other types can be observed. This demonstrates that significantly smaller training sets can still be sufficient to obtain well-performing models. It has to be noted that similar trends were observed for the baseline DNN model with test data from types earthquake, flood and hurricane (Wiegmann et al. 2020a). For $N < 250$, GP F_1-scores dramatically decrease. One reason for this might be the fact, that binary GP model optimization is initialized by so called inducing variables representing the latent space points to be optimized. In order to ensure computational tractability, we used 100 normally distributed random sample points covering the latent space, which might not be an optimal choice - especially in case of few training samples.

Table 2. Average F_1-scores of binary GP models trained with N random samples, corresponding baseline DNN scores with $N = 3000$ training samples, and average false positive rates (FPR) from classifying 5 million unrelated tweet texts. For each test event type, L labels from E events are available.

Test Type			Training Sample Size N									
			Binary GP Models									DNN
	L	E	50	100	250	500	750	1,000	1,500	2,000	3,000	3,000
Biological	6,106	2	0.523	0.672	0.910	0.911	0.911	0.920	0.921	0.915	0.930	**0.974**
Flood	14,210	9	0.509	0.662	0.870	0.874	0.873	0.888	0.884	0.891	**0.895**	0.840
Wildfire	4,820	3	0.516	0.682	0.916	0.919	0.924	0.922	0.932	0.932	0.938	**0.949**
Hurricane	48,922	9	0.519	0.670	0.890	0.898	0.899	0.911	0.913	0.914	**0.918**	0.858
Earthquake	27,034	11	0.519	0.673	0.903	0.906	0.919	0.923	0.928	0.931	**0.933**	0.915
Tornado	11,204	2	0.520	0.685	0.932	0.935	0.937	0.939	0.945	0.948	**0.957**	-
Transport	3,850	4	0.525	0.697	0.931	0.937	0.937	0.931	0.942	0.949	0.953	**0.954**
Societal	11,206	2	0.518	0.663	0.871	0.887	0.878	0.873	0.877	0.862	0.879	**0.899**
Industrial	10,166	4	0.518	0.693	0.938	0.941	0.939	0.944	0.950	0.950	0.957	**0.964**
All	137,518	46	0.519	0.675	0.902	0.909	0.908	0.912	0.917	0.915	**0.925**	0.883
Tranquil (FPR)	5 Million	-	0.475	0.310	0.079	0.076	0.069	0.067	0.059	0.059	0.048	0.047

Average FPRs of GP and DNN models are similar for $N = 3,000$. Even though the FPR of GP models is nearly doubled for $N = 250$, this might still be a suitable model to automatically identify more representative tweets reporting on a specific event. The obtained results indicate that DNNs and GPs offer advantages in different surrounding conditions. Further in-depth experiments are required to gain more insights here. The question of how GP model variances can help to evaluate prediction uncertainties will also be part of future work.

One-class GP Models

The main challenge in OCC is that the GP hyperparameters cannot be learned from labeled data. Since we therefore do not use fully connected layers, our one-class model has only two hyperparameters: the kernel lengthscale l and variance σ^2 (even though one could also use an independent pair of parameters for each GP dimension). To gain first insights regarding the sensitivity of these parameters, we focus on the filtering task involving data covering all event types. In each CV iteration, a parameter grid search is performed. According to preliminary experiments, the parameters were defined as $\log(l) \in \{-1.5, \ldots, 1.0\}$ and $\log(\sigma^2) \in \{-3.5, \ldots, 1.0\}$. With a step size of $\delta = 0.1$, a total number of 1,125 configurations are tested each CV. Since off-topic samples are not required, only $N/2$ positive samples are used for training. The best identified models are additionally applied to the tranquil data to investigate the resulting FPR. The best obtained average results over all CVs and the corresponding model parameters are listed in Table 3.

Table 3. Best average 10-fold CV F_1-scores obtained with one-class GP models applied to independent test data from all event types. The scores are compared to a baseline DNN model ($N = 3.000$). For each N, the best parameters l and σ^2 are found in a grid search and then used to filter the tranquil set (FPR).

	Training Sample Size N									
	One-class GP Models									DNN
	25	50	125	250	375	500	750	1,000	1,500	3,000
F_1	0.854	0.873	0.881	0.879	**0.887**	0.881	0.885	0.885	0.882	0.883
FPR	0.119	0.110	0.114	0.110	0.103	0.114	0.111	0.116	0.125	**0.047**
	Best GP Model Parameters									
l	0.606	0.549	0.549	0.606	0.606	0.549	0.549	0.496	0.496	
σ^2	0.819	0.905	0.223	0.055	0.037	0.050	0.033	0.055	0.037	

For a broad range of training sample sizes $50 \leq N \leq 1,500$, the obtained F_1-scores are similar or even slightly better compared to the binary baseline model ($N = 3,000$). In comparison to experiment 1 (all events), the best obtained one-class F_1-score of 0.887 is around 0.04 higher compared to the best DNN model and around 0.05 lower compared to the best binary GP model. However, a better performance of GP models with additional fully connected layers comes at the cost of more required training data. In the binary setting, the maximum F_1 is obtained with $N = 3,000$ training samples, where $N = 375$ samples were sufficient for the one-class model. No significant

F_1 drops can be observed in case of $25 \leq N \leq 375$ training samples. Since our binary models tend to provide rather random predictions in these cases, a model without fully connected layers turns out to be more suitable here. A further reason for performance differences might be the fact that another GP implementation (GPflow (Matthews et al. 2017)) is used, avoiding the need of inducing variables.

For all training sample sizes, the FPR is approximately three times higher compared to the FPR of the binary baseline model, reflecting the lack of adjusted latent space decision boundaries estimated by incorporating negative training samples. This might be acceptable, if such a model could be directly applied for live stream overload reduction without the need of iteratively searching parameters and only requiring up to $N = 50$ representative samples. However, appropriate values for the lengthscale and variance are unknown *a priori*. Maximum differences $\Delta l = 0.11$ and $\Delta\sigma^2 = 0.87$ of the found optimal parameters indicate a rather narrow search space for the investigated task. Average values $\bar{l} = 0.56$ and $\bar{\sigma^2} = 0.25$ may serve as a reasonable choice for initial parameters. Strong observed F_1 peaks with respect to the lengthscale for all tested one-class models demonstrate that the log-scale step-size $\delta = 0.1$ is not optimal and should be decreased.

CONCLUSION AND OUTLOOK

In this work, first experimental results using Gaussian processes for Twitter stream overload reduction are presented. The basic idea is to combine current deep learning-based feature representations with the Bayesian framework of GPs in order to approach Twitter overload reduction in binary and one-class settings. The investigated models are intended to identify any crisis-related microblog texts from unfiltered Twitter streams.

Our binary GP classification model, in which the two-node output layer of a standard DNN is replaced by a GP, is able to outperform the baseline DNN in case training data from multiple events are available. In case training data from only few events are available, the baseline DNN shows a bit better generalization capability. With $N \leq 250$ labeled samples, training of the binary GP model does not lead to satisfactory cross-event test results. One reason for this might be the high number of fully connected layer weights ($N_w = 67,728$) that need to be estimated together with the GP parameters. A further reason could be the sub-optimal initialization of inducing variables.

Our one-class model comprises a pre-trained sentence embedding and a single kernel regression GP and therefore only has two trainable parameters. This is intended to reduce the barriers for model application in realistic scenarios. Another explanation for this design is that end-to-end training is simply not possible in the OCC case. Compared to the baseline DNN, similar F_1-scores can be observed for all investigated training set sizes $25 \leq N \leq 1,500$, indicating the great potential of OCC GPs. However, reduced efforts in training data sampling and training efforts come at the cost of nearly tripled false positive rates of up to 0.12. Furthermore, it has to be noted that also binary DNNs are able to provide quite good results in case of few training samples.

According to our experimental results, the main advantages of one-class GPs can be summarized as follows: (1) OCC GPs don't require negative training samples and therefore the problem of extremely unbalanced class distributions and the related task of sampling representative samples is circumvented; (2) Our investigated OCC GPs do not require DNN parameter optimization and are therefore potentially well suited for immediate social media situational monitoring in case of completely new event or topic types.

In conjunction with the challenge of finding optimal one-class GP model parameters, further research in several directions is required. In future work, more experiments involving baseline models for all tested configurations and training sample sizes N will be conducted. Besides the usual evaluation metrics, also qualitative investigations are planned. This will help to better understand which model to prefer in case of specific surrounding conditions. The benefits of the GP model prediction variances will also be investigated, since this information is a potential indicator for unreliable model outputs and may therefore be helpful to detect uncertain predictions as well as for a guided collection of new training data. Furthermore, the impact of utilizing different tailored text embeddings, like BERTweet (Nguyen et al. 2020) or CrisisBERT (Liu et al. 2020), will be investigated.

The fact that one-class GP models require manual interaction can be seen as an advantage for practical applications, like flexible social media-based situational monitoring. A one-class GP model along with few positive samples could be used to obtain initial stream filtering results, for example to identify crisis-related tweets or to find tweets that contain eyewitness reports. We therefore plan to implement and test an interactive dashboard that enables users without deep machine learning and programming experiences to train, apply and interactively adjust GP models. In order to gain more insights regarding different application scenarios, additional labeled data sets, like HumanAID (Alam et al. 2021) or tweets representing eyewitness reports (Zahra et al. 2020), will be incorporated in future experiments.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.

Alam, F., Qazi, U., Imran, M., and Ofli, F. (2021). "HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*. Ed. by C. Budak, M. Cha, D. Quercia, and L. Xie. AAAI Press, pp. 933–942.

Beck, D. E. (2017). "Gaussian Processes for Text Regression". PhD thesis. University of Sheffield.

Bradshaw, J., Matthews, A. G. D. G., and Ghahramani, Z. (2017). "Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks". In: *arXiv: Machine Learning*.

Burel, G. and Alani, H. (May 2018). "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, p. 12.

Burges, C. and Burges, C. J. (Jan. 1998). "A Tutorial on Support Vector Machines for Pattern Recognition". In: *Data Mining and Knowledge Discovery* 2, pp. 121–167.

Camps, S., Houben, T., Fontanarosa, D., Edwards, C., Antico, M., Dunnhofer, M., Martens, E., Baeza, J., Vanneste, B., Limbergen, E. van, et al. (2018). "One-class Gaussian process regressor for quality assessment of transperineal ultrasound images". In: *Proceedings of the 1st International Conference on Medical Imaging with Deep Learning 2018*. Ed. by I. Isgum, C. Sanchez, and G. Litjens. The Netherlands: Medical Imaging with Deep Learning Conference Committee, pp. 1–10.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). "Universal Sentence Encoder". In: *CoRR* abs/1803.11175. arXiv: 1803.11175.

Cohn, T. and Specia, L. (Aug. 2013). "Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 32–42.

Damianou, A. and Lawrence, N. D. (Apr. 2013). "Deep Gaussian Processes". In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. Ed. by C. M. Carvalho and P. Ravikumar. Vol. 31. Proceedings of Machine Learning Research. Scottsdale, Arizona, USA: PMLR, pp. 207–215.

Deborah, S. A., Rajendram, S. M., and Mirnalinee, T. (Aug. 2017). "SSN_MLRG1 at SemEval-2017 Task 4: Sentiment Analysis in Twitter Using Multi-Kernel Gaussian Process Classifier". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 709–712.

Dutordoir, V., Salimbeni, H., Hambro, E., McLeod, J., Leibfried, F., Artemev, A., Wilk, M. van der, Deisenroth, M. P., Hensman, J., and John, S. (2021). "GPflux: A library for Deep Gaussian Processes". In: *arXiv:2104.05674*.

Fathi, R., Thom, D., Koch, S., Ertl, T., and Fiedrich, F. (2020). "VOST: A case study in voluntary digital participation for collaborative emergency management". In: *Information Processing & Management* 57.4, p. 102174.

Hiltz, S. R. and Plotnick, L. (2013). "Dealing with information overload when using social media for emergency management: Emerging solutions". In: *ISCRAM*.

Jayashree, P. and Srijith, P. K. (May 2020). "Evaluation of Deep Gaussian Processes for Text Classification". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1485–1491.

Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. (June 2010). "Gaussian Processes for Object Categorization". In: *International Journal of Computer Vision* 88.2, pp. 169–188.

Kaufhold, M.-A., Bayer, M., and Reuter, C. (2020). "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning". In: *Information Processing & Management* 57.1.

Kemmler, M., Rodner, E., Wacker, E.-S., and Denzler, J. (Dec. 2013). "One-Class Classification with Gaussian Processes". In: *Pattern Recogn.* 46.12, pp. 3507–3518.

Kersten, J., Kruspe, A., Wiegmann, M., and Klan, F. (2019). "Robust Filtering of Crisis-related Tweets". In: *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Valencia, Spain.

Kruspe, A., Kersten, J., and Klan, F. (2021). "Review article: Detection of actionable tweets in crisis events". In: *Natural Hazards and Earth System Sciences* 21.6, pp. 1825–1845.

Kruspe, A., Kersten, J., and Klan, F. (2019). "Detecting event-related tweets by example using few-shot models". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, Valencia, Spain, May 19-22, 2019*.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). "When Gaussian Process Meets Big Data: A Review of Scalable GPs". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.11, pp. 4405–4423.

Lukasik, M., Bontcheva, K., Cohn, T., Zubiaga, A., Liakata, M., and Procter, R. (Feb. 2019). "Gaussian Processes for Rumour Stance Classification in Social Media". In: *ACM Trans. Inf. Syst.* 37.2.

Lukasik, M., Cohn, T., and Bontcheva, K. (Sept. 2015). "Classifying Tweet Level Judgements of Rumours in Social Media". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2590–2595.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (Apr. 2017). "GPflow: A Gaussian process library using TensorFlow". In: *Journal of Machine Learning Research* 18.40, pp. 1–6.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (July 2019). "A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets". In: *International Journal of Information Systems for Crisis Response and Management* 11.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (Oct. 2020). "BERTweet: A pre-trained language model for English Tweets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 9–14.

Ober, S. W., Rasmussen, C. E., and Wilk, M. van der (2021). *The Promises and Pitfalls of Deep Kernel Learning*. arXiv: 2102.12108 [stat.ML].

Plotnick, L. and Hiltz, S. R. (2016). "Barriers to Use of Social Media by Emergency Managers". In: *Journal of Homeland Security and Emergency Management* 13.

Preoţiuc-Pietro, D. and Cohn, T. (Oct. 2013). "A temporal model of text periodicities using Gaussian Processes". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 977–988.

Rao, R., Plotnick, L., and Hiltz, R. (2017). "Supporting the use of social media by emergency managers: Software tools to overcome information overload". In: *Proceedings of the 50th Hawaii international conference on system sciences*.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Reuter, C. and Kaufhold, M.-A. (2018). "Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics". In: *Journal of Contingencies and Crisis Management (JCCM)* 26.1, pp. 41–57.

Roman, I., Mendiburu, A., Santana, R., and Lozano, J. A. (2019). "Sentiment Analysis with Genetically Evolved Gaussian Kernels". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '19. Prague, Czech Republic: Association for Computing Machinery, pp. 1328–1337.

Snyder, L. S., Lin, Y., Karimzadeh, M., Goldwasser, D., and Ebert, D. S. (2020). "Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness". In: *IEEE Trans. Vis. Comput. Graph.* 26.1, pp. 558–568.

Stieglitz, S., Mirbabaie, M., Fromm, J., and Melzer, S. (2018). "The Adoption of Social Media Analytics for Crisis Management – Challenges and Opportunities". In: *Twenty-Sixth Eur. Conf. Inf. Syst. (ECIS2018)*.

van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V., and Hensman, J. (2020). "A Framework for Interdomain and Multioutput Gaussian Processes". In: *arXiv:2003.01115*.

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (2020a). "Analysis of Filtering Models for Disaster-Related Tweets". In: *Proceedings of the 17th ISCRAM, May 24-27*. ISCRAM. Blacksburg, Virginia (USA).

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (Mar. 2020b). *Disaster Tweet Corpus 2020*. Version 1.0.0. Zenodo.

Wiegmann, M., Kersten, J., Senaratne, H., Potthast, M., Klan, F., and Stein, B. (2021). "Opportunities and risks of disaster data from social media: a systematic review of incident information". In: *Natural Hazards and Earth System Sciences* 21.5, pp. 1431–1444.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (Sept. 2016). "Deep Kernel Learning". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Gretton and C. C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, pp. 370–378.

Zahra, K., Imran, M., and Ostermann, F. O. (2020). "Automatic identification of eyewitness messages on twitter during disasters". In: *Information Processing & Management* 57.1, p. 102107.

Zhang, C., Fan, C., Yao, W., Hu, X., and Mostafavi, A. (2019). "Social media for intelligent public information and warning in disasters: An interdisciplinary review". In: *International Journal of Information Management* 49, pp. 190–207.