Micropost Incident Detection

Supervisor: Prof. Dr. Michael Granitzer

Advisor: Lorenz Wendlinger

Scope: Bachelor/Master

Abstract

Social media postings can be a great early warning system for hyper-local as well as larger-scale incidents and crises. The large amount of data generated on social media platforms makes manual examination of these postings infeasible. Even automated fact-extraction systems may require some form of filtering to keep up.

For this purpose, automatic incident detection is necessary in order to reject irrelevant postings early on. We consider a posting relevant to a type of incident, if it contains one or multiple facts about an incident of the respective type. The detection can therefore be formulated as a multi-label or multi-class classification problem.

One approach to solving this are word embedding based Convolutional Neural Networks. The curse of dimensionality resulting from the large categorical vocabulary of natural language can be solved by projecting words into a dense embedding space. These embeddings are learned on large corpora, and can largely preserve semantic similarities between tokens. Consequently, we can use this to obtain a dense variable-length real-valued representation of a posting. These can then be processed with 1D convolutions and global pooling to obtain a shift-invariant fixed-length representation, that an MLP can learn relevance from (c.f. [1, 2]).

As an extension of this model, the C-BiGRU from [3] adds a bi-directional Gated Recurrent Unit before the global pooling to capture long-term dependencies in the input.

This kind of convolution is based on the neighborhood of tokens in a sentence represented as a 1D sequence. However, there are other dependencies between tokens that offer a different neighborhood definition. This transforms the structure of a sentence from a simple sequence into a directed acyclic graph (DAG). Such dependencies can be extracted automatically through dependency parsing, e.g. [6, 7]. If this works well enough for social media postings, graph convolutions can be used on the dependency graph to obtain a fixed length representation, similar to the 1D convolutions described above.

Some incident detection datasets have been developed based on Twitter postings. *MMoveT15*[4] is a collection of tweets annotated regarding their relevance to migration movements. *MMoveT15+* is an extended version annotated based on a more fine-grained relevance definition. *Incident related Twitter*[5] contains mentions of fires, shootings and crashes collected from multiple english-speaking cities.

As they only give little context and frequently contain misspellings or colloquial language, social media postings are challenging for such approaches. It is therefore interesting how the choice of word embedding method and corpus influences results. Their multilinguality also requires certain adaptations, such as using (models trained on) multilingual corpora or multiple models.

Task

- Analyze datasets
- Develop baselines for relevance classification (Entry Task)

- Adapt the CNN model of [1, 2] to incident detection
- Adapt the C-BiGRU model of [3] to incident detection
- Compare different Word Embeddings (methods and base corpora)
- Conduct hyper-parameter study (network architecture, training schema)
- (optional) Use dependency parsing and graph convolutions to learn from a more salient token neighborhood
- (optional) Compare different methods for processing hashtags (e.g. remove, treat as regular text, process separately)
- (optional) Experiment with additional features (e.g. POS tags, sentiment scores)

Recommended Competences

- Strong development skills in Python
- Familiarity with the NLTK or related NLP libraries
- Experience with DL libraries (`PyTorch`, `Tensorflow` or `Keras`)
- Good grades in relevant courses (e.g. text mining, text mining project, data science seminar, data science lab)

Entry Task

Implementation using `scikit-learn`, delivery as a standalone `jupyter` notebook

- Construct a relevance classifier from a tfidf-weighted BoW tokenizer and an SVM
- Optimize the n-gram range of the BoW tokenizer on the train partition of [4] with 5-fold CV
- Train model on the whole train partition
- Use platt-scaling to obtain prediction probabilties for the test partition
- Report 10 postings with the highest positive probability

Application

- Apply before starting with the entry task, but you can ask questions about the entry task in your application.
- Candidates who apply for the topic should provide a transcripts of records.
- Start the entry task as soon as you get a positive response to your application.
- The Entry Task needs to be finished within 7 days.
- Contact: lorenz.wendlinger@uni-passau.de

---

[1] Rakhlin, A. “Convolutional neural networks for sentence classification.” arXiv (2014).

[2] Severyn, Aliaksei, and Alessandro Moschitti. “Unitn: Training deep convolutional neural network for twitter sentiment classification.” Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). 2015.

[3] Mitrović, Jelena, Bastian Birkeneder, and Michael Granitzer. “nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection.” Proceedings of the 13th International Workshop on Semantic Evaluation. 2019.

[4] Urchs, Stefanie, Lorenz Wendlinger, Jelena Mitrović, and Michael Granitzer. “MMoveT15: A Twitter Dataset for Extracting and Analysing Migration-Movement Data of the European Migration

Crisis 2015." In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 146-149. IEEE, 2019.

[5] Schulz, Axel, Christian Guckelsberger, and Frederik Janssen. "Semantic abstraction for generalization of tweet Classification." Semantic Web 1.2014 (2014): 1-5.

[6] https://spacy.io/api/dependencyparser

[7] https://nlp.stanford.edu/software/nndep.html