Natural Language Processing and Information Extraction: Project

1 Introduction

The project exercise described in this document will allow you to get acquainted with all steps of solving a complex NLP task, starting with data preprocessing and simple baseline solutions in the first weeks of the course and moving towards more complex approaches by the second half of the semester. Throughout the text, various deadlines are referred to. All deadlines and what is due by these deadlines are summarized in Section 8.

2 Overview

In the first weeks of the semester you are asked to **form groups of 4 and choose a task** from Section 5, based on which one or two course instructors will become your group's **mentor(s)**, they will support you throughout the semester and evaluate each of your submissions. Your group will be added to a GitHub repository for pushing your submissions, details are in Section 4.

Each task will involve the use of 2-3 datasets, these have to be properly preprocessed, often using some of the methods described in **Lecture 1** (Text processing). The next step is the implementation of some simple, standard solutions, these are often called **baselines**. In **Lecture 2** (Text classification) some simple, task-independent approaches will be introduced, and **Milestone 1** is the implementation of at least one simple baseline for your chosen task on at least one dataset, plus a very brief analysis of its strengths and weaknesses (not more than 1-3 paragraphs of text). This should be submitted by the end of **Week 6**, in the form of clean and documented Python code pushed to your project repository. Deep learning approaches in NLP will be introduced in **Lectures 3-6**, after this you will be able to train and evaluate a neural network architecture appropriate for your chosen task, this is **Milestone 2** and should be submitted by the end of **Week 9**, also in the form of clean and documented Python code.

In the second half of the semester the lectures will introduce approaches to modeling linguistic structure and meaning, then provide an overview of approaches to some of the most common tasks in NLP, any of which may be applicable to your chosen topic. Groups are expected to conceive and implement approaches that go beyond the standard baselines implemented in the first half of the semester. The value of these solutions may come not only from superior quantitative performance, but also from better explainability, from broader applicability (e.g. different domains, less data), simplicity, etc. You are encouraged to approach your mentor to discuss your ideas and get feedback. The final results must be communicated in both a 30 minute presentation (where all team members present their own contributions) and a 2-page management summary, and should be accompanied by a clean and readable software repository. For general and topic-specific instructions, see Sections 3 and 5, respectively. In the last week of the course each group will present their project, and should have submitted their software and report, all of which will be evaluated separately. Presentations will take place on **Week 13**, the deadline for the final submission is on Week 14, see Section 6 for detailed instructions and Section 7 on evaluation principles.

3 General instructions

3.1 Goals

The topic descriptions in Section 5 provide many pointers and ideas for getting started, and indicate some challenges and questions that you can work on. You are not expected to address more than 1-2 of the challenges and questions listed, but the value of your project comes from your contributions to these (the implementation of standard methods with existing datasets can only satisfy Milestones 1 and 2). Quantitative performance of a solution is only one indicator of its value, based on the topic and the nature of your solution you may also need to consider aspects such as complexity, explainability, sustainability, risk of unintended bias, applicability (to multiple domains, datasets, or languages), etc.

3.2 Datasets and languages

Each topic description makes some recommendations on datasets, but you are encouraged to find additional resources. Using datasets in languages other than English or German that are understood by members of your group is encouraged, and so is working on more than one language in the project. If you choose a language for which datasets are already available, consider using at least two of them in the project. You may also choose a language with no datasets, in this case your main challenge will be to find possible ways to bootstrap a solution and/or a dataset.

3.3 Evaluation

Proper evaluation of methods, including your own, both quantitative (e.g. precision and recall) and qualitative (e.g. looking at the data), is essential. For some tasks and some datasets you cannot assume that higher figures mean better solutions. Some manual analysis of a system's output is usually necessary to understand its strengths and limitations. Topic descriptions may indicate task-specific challenges of evaluation.

4 Technical

4.1 Version control

After teams registered for topics, GitHub Classroom repositories will be created for each team by the instructors and team members will be invited to join. Teams should then push their solutions to this repository.

4.2 Coding guidelines

Your solution should be implemented in **Python 3.7** or higher and should generally conform to **PEP8** guidelines. You should also observe **clean code** principles.

5 Topics

You may select one of the following topics or come up with your own topic (in the latter case, check with the exercise coordinator). Each topic has a limit on the number of groups that can select it (use TUWEL for registering groups and topics). Read carefully both the general instructions in Section 3 and those specific to your chosen task below. The instructor listed for your chosen topic will be your point of contact in case of questions, you are encouraged to consult them. E-mail addresses are listed below:

Alaa El-Ebshihy `alaa.el-ebshihy@tuwien.ac.at`

Kinga Gémes kinga.gemes@tuwien.ac.at

Ádám Kovács adam.kovacs@tuwien.ac.at

Annisa Maulida Ningtyas annisa.ningtyas@student.tuwien.ac.at

Gábor Recski gabor.recski@tuwien.ac.at

5.1 Topic 1: Detection of offensive text in social media

Instructor Kinga Gémes, Gábor Recski

Overview The goal of this task is the classification of short utterances on social media (e.g. tweets, facebook comments, etc.) to determine whether they are offensive. Fine-grained classification may involve classes of offensive text such as hate speech, abuse, profanity, etc.

Resources Several datasets from the last few years are freely available online for both English (Zampieri et al., 2019; Zampieri et al., 2020; Mandl et al., 2019; Mandl et al., 2020; Mandl et al., 2021) and German (Graf and Salini, 2019; Struß et al., 2019; Risch et al., 2021; Mandl et al., 2019; Mandl et al., 2020). Some shared tasks (Zampieri et al., 2020; Mandl et al., 2019; Mandl et al., 2020; Mandl et al., 2021) offer datasets in several languages.

If you wish to use a different dataset, or focus on a more specialized offensive text detection problem, such as sexism or xenophobia (Basile et al., 2019; Kumar et al., 2020), please contact your instructor.

Questions and challenges

- The definition of what is offensive and the consistency of annotation varies across datasets, this is one of the main challenges of the task. Is there really an objective way to define the task or evaluate a system?
- Users of this technology will often have specific examples of what text they want to see classified as offensive or non-offensive. How could your solution accommodate such needs?
- How can you make your classifier's decisions explainable to users? What are the limitations of the explanations you can provide?

5.2 Topic 2: Relation Extraction

Instructor Ádám Kovács, Gábor Recski

Overview Relation extraction (RE) is the task of extracting semantic relationships between entities from a text. These relationships occur between two or more entities and are defined by certain semantic categories (e.g. Destination, Component, Employed by, Founded by, etc..). Entities usually fall into certain types (e.g. Organization, Person, Drug type, Location, etc..). The task is to build a classifier that learns to predict the relationship between entities. Let's have an example sentence with two entities as relation candidates:

> **Elevation Partners**, the $1.9 billion private equity group that was founded by **Roger McNamee**.

Typically in RE tasks, two entities (in our case, *Elevation Partners* and *Roger McNamee*) and usually their types (COMPANY, PERSON) are given in a context (e.g. in a sentence), and the task is to classify the *relation* that the two entity holds (if there is any). For this example, the correct label would be *founded_by*.

Resources

- Generic relation extraction datasets e.g. the Semeval 2010 dataset (Hendrickx et al., 2010), or the TACRED (Zhang et al., 2017).
- Domain specific relation extraction on medical data.
  - The CrowdTruth dataset (Dumitrache, Aroyo, and Welty, 2018) and the FoodDisease dataset (Cenikj, Eftimov, and Koroušić Seljak, 2021). In both task the *cause* or *treat* relation should be classified between drugs and foods.
  - Other medical relation extraction resources from the BLUE benchmark: the `DDI` (Herrero-Zazo et al., 2013), `ChemProt` (Taboureau et al., 2011) or the `i2b2 2010 shared task` (Uzuner et al., 2011) dataset

Questions and challenges

- RE differs from classical classification tasks in that information about the relation candidates (the two entities in question) also needs to be modeled. How would you construct such a machine learning model for a RE task?
- How would you leverage graphs (e.g. Universal Dependency trees) into your solution (idea: use paths between the entities as features)?
- **Advanced question:** The CrowdTruth and the FoodDisease datasets contain the same labels and similar entity types. How do modern neural based models (e.g. BERT (Devlin et al., 2018)) transfer their knowledge between the datasets? Do rule-based models transfer better?

5.3 Topic 3: Natural Language Inference

Instructor Ádám Kovács, Gábor Recski

Overview Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*). The *premise* can entail, contradict or be neutral to the hypothesis. Entailment is meant when the premise would infer that the hypothesis is true (Dagan, Glickman, and Magnini, 2006). The task is to classify a set of *premise-hypothesis* pairs into a category (entailment, contradiction, neutral).

An example *premise-hypothesis* pair for the entailment class would be:

premise - A young family enjoys feeling ocean waves lap at their feet.

hypothesis - A family is at the beach

And for contradiction:

premise - There is no man wearing a black helmet and pushing a bicycle

hypothesis - One man is wearing a black helmet and pushing a bicycle

For a detailed overview, see this seminar: https://nlp.ec.tuwien.ac.at/seminar/sessions/20210302/nli.pdf

Resources

- Smaller datasets like SICK (Marelli et al., 2014), an english corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena).
- More recent datasets with hundred thousand examples, like SNLI (Bowman et al., 2015) or Multi-NLI (Williams, Nangia, and Bowman, 2018)

Questions and challenges

- In NLI, instead of just classifying a sentence, the task is to classify a *pair* sentence into one of the categories. How would you formulate and implement a machine learning model to solve this problem?
- Both the SNLI and the Multi-NLI datasets contain thousands of labeled examples. There is a re-occurring problem in these datasets in that the data leaves clues about the labels, enabling neural models to learn these artifacts instead of the real problem. Could you find some of these clues in the data and use them as features for your model (hint: look at the paper (Gururangan et al., 2018))?
- **Advanced question**: (Talman and Chatzikyriakidis, 2019) tested the generalization power of neural models on multiple datasets, concluding that models that incorporate knowledge-base into their prediction does better on other datasets. Additional experiments could be made with more modern BERT (Devlin et al., 2018) based approaches and also rule-based systems. Does a more traditional approach transfer it's knowledge of NLI better?

5.4 Topic 4: Attribute extraction from building regulations

Instructor Gábor Recski

Overview As part of the BRISE project, text documents of the Zoning Plan (Flächenwidmungsplan) of the City of Vienna have been annotated for the attributes they regulate. For example, the sentence *Der oberste Abschluss der zur Errichtung gelangenden Gebäude darf 15 m nicht überschreiten.* is annotated with the attribute `GebaeudeHoeheMax`. A small sample is also annotated for the values of attributes and the rule structure, e.g. that the value of `GebaeudeHoeheMax` is `15m` and the modality of the rule is `obligation`. The main task is the classification of sentences based on which attributes they mention. Possible next steps involve choosing some frequent attributes and also extracting their values from the text.

Resources

- Small annotated samples are publicly available in the brise-plandok repository, a larger set of annotated documents (attributes only) is available from the instructor upon request.
- The annotation guidelines are available here
- A simple rule-based extraction system is already available here.

Questions and challenges

- Analyze the performance of your classifiers: which of the most frequent labels are also hard to detect?
- Compare your ML-based solutions to the existing rule-based system for the most frequent labels. What are the advantages of each? Which one would you choose? Would you have a way of combining them?
- For some attributes it is straightforward to also get the values, using simple patterns. Which ones are more difficult? How would you approach them?

5.5 Topic 5: Argumentative Zoning of scientific articles

Instructor Alaa El-Ebshihy

Overview Argumentative Zoning (AZ) is the analysis of the argumentative and rhetorical structure of a scientific paper (Teufel et al., 1999). The basic idea of AZ is to assign each sentence in the scientific article to a specific category (known as zone). Each zone represents one of the article's component (e.g. the hypothesis, the background, the method, .. etc). AZ was first defined by Simon Teufel (Teufel et al., 1999; Teufel and Moens, 2002) where she defined a schema of 7 zones. A more fine grained schema with 15 categories was defined in a later work (Teufel, Siddharthan, and Batchelor, 2009). A direct application to AZ is the generation of user-tailored summaries for scientific articles using the defined regions in the article text.

The goal of this task is to identify only the regions in the scientific article that represent **the work done by its authors** and represent the following aspects: the claim of the authors, the methodology, the results of the experiments and the conclusions.

As an illustrative example for the task, let's have a scientific paper P05-1053 in which we find sentences belonging to the work of authors. The sentences are then classified using simplified schema of 4 categories: *Claim, Method, Result and Conclusion*. Some of these sentences:

Claim - This paper focuses on the ACE RDC task and employs diverse lexical, syntactic and semantic knowledge in feature-based relation extraction using Support Vector Machines (SVMs).

Method - Semantic information from various resources, such as WordNet, is used to classify important words into different semantic lists according to their indicating relationships.

Result - The experiment result also shows that our feature-based approach outperforms the tree kernel-based approaches by more than 20 F-measure on the extraction of 5 ACE relation types.

Conclusion - Last, effective ways need to be explored to incorporate information embedded in the full parse trees

Resources

- Some full annotated articles datasets are available. This includes the original AZ dataset (Teufel et al., 1999; Teufel and Moens, 2002), the AZ chemistry dataset with the extended schema (Teufel, Siddharthan, and Batchelor, 2009) and the AZ Biomedical articles dataset (Guo et al., 2013).
- Some other datasets available which includes the annotation of the abstracts of the articles only. For example the SciArg dataset (Accuosto, Neves, and Saggion, 2021) and the AZ Biomedical abstract dataset (Guo et al., 2010).
- For manual evaluation purpose, a selected dataset of scientific articles and online annotation tool are available. Please contact your instructor.
- Guidelines for defining the zones is also available for the SciArg (Accuosto, Neves, and Saggion, 2021) dataset and guidelines prepared by the instructor.
- For parsing scientific article, GROBID [1] can be used and available as an API for direct usage. Please contact the instructor for the API.

Questions and challenges

- Due to the complexity of the annotation of scientific articles, existing fully annotated training data is quiet small. Would increasing the size of a dataset lead to a better performance?

[1] https://grobid.readthedocs.io/en/latest/Introduction/

- Writing style in scientific articles differs from one domain to another (for example, the writing style of articles from Biomedical domain differs from that of the Computer Science domain), therefor most of available models are trained on domain specific datasets. What is the effect of applying a trained model on a specific domain to a different one?
- The implemented model may introduce some limitation (e.g. due to the training dataset size or how old is the training data or the domain). How can we do manual evaluation to identify if any of the limitations exists?

5.6 Topic 6: Medical Concept Normalization

Instructor Annisa Maulida Ningtyas

Overview Medical Concept Normalization (MCN) for user-generated text aims to map a health condition described in colloquial language to standard medical terminology such as "Systematized Nomenclature of Medicine - Clinical Terms" (SNOMED-CT). For instance, the colloquial expression "can't fall asleep all night" was mapped onto the medical term insomnia. MCN is important for bridging the gap between laypeople's informal medical phrases and formal medical terminologies that can be used for further analysis.

This task aims to normalize colloquial medical phrases into formal medical phrases from the medical knowledge base.

Resources

- CADEC: CSIRO Adverse Drug Event Corpus (CADEC) includes user-generated medical reviews related to Diclofenac and Lipitor (Karimi et al., 2015). The manually identified health-related mentions are mapped to target concepts in SNOMED-CT vocabulary. The dataset consists of 6,754 mentions, each of which corresponds to one of the 1029 SNOMED-CT codes. Tutubalina et al. (Tutubalina et al., 2018) create custom folds with minimal overlap because the random folds of the CADEC dataset created by Limsopatham and Collier (Limsopatham and Collier, 2016) have significant overlap between train and test instances.
- PsyTAR: Psychiatric Treatment Adverse Reactions (PsyTAR) corpus includes psychiatric drug reviews obtained from AskaPatient (Zolnoori et al., 2019). Zolnoori et al. (Zolnoori et al., 2019) manually identify 6556 health-related mentions and map them to one of 618 SNOMED-CT codes. Due to significant overlap between train and test sets of random folds released by Zoolnori et al. (Zolnoori et al., 2019), Miftahutdinov and Tutubalina (Miftahutdinov and Tutubalina, 2019) create custom folds of this dataset with minimum overlap.

Questions and challenges

- The majority of MCN research treats the problem as a multi-classification task. However, SNOMED-CT contains around 350,000 concepts, and the current datasets do not cover all of them. This may result in unseen colloquial phrases with uncovered SNOMED-CT concepts being mapped into the incorrect medical concept. Please demonstrate this assumption!
- Based on your observation in the first challenges, how can we broaden the concept coverage of the current datasets?

6 Submission, report, and final presentation

Submission: The deadline for pushing your final commits to the project repository is the 13th of January 2022 at 23:55 CET.

Report: Your submission must be accompanied by a **2-page PDF document** that presents a summary of your solution — this is a management summary, so it should be written in a way that is easy to understand by top management, not NLP colleagues. The summary should contain an overview of the task, the challenges you faced, the external resources you used, the solution you implemented, and a short discussion of where you now stand.

Final Presentation: On the 14th of January 2022, **each group will present the main results of their work to all other groups working on the same topic.** The format is **20 minutes of presentation and 10 minutes of discussion** — we will be very strict with the timing, and stop the presentation at the 20 minute mark. **Each team member must present their own contributions to their project, so that they can be evaluated individually.** The presentation should be aimed at NLP colleagues, so highlight which approaches and techniques you used, which data you used, and the insights obtained. By the 13th of January you should also push the presentation slides to your project repository. Everybody must attend all presentations for their topic.

7 Evaluation

The final mark will be based on the submitted code and report as well as the presentation. Milestone 1 and Milestone 2 must each be completed with a minimum score of 35% by their respective deadlines to pass the course.

The final mark is calculated from the following components:

- 15% for Milestone 1
- 15% for Milestone 2
- 50% for the final solution
- 10% for the presentation
- 10% for the management summary

Note that about 50 hours per person is foreseen for this exercise, around two-thirds of the time foreseen for the course (75 hours). This means that everyone should work more than a standard (40 hour) week on this exercise, so four weeks effort for a group of four. The evaluation will be based on the expectation of a manager (an NLP expert) assigning such a task to a team of four junior NLP engineers for a week. Note that this expectation is not met by submitting an overly long Jupyter notebook — you need to demonstrate that:

- You have approached the analysis in a logical and structured way.
- You are aware of the solutions already available for the problem, and show how your solution builds on them (note that you don't need a comprehensive state-of-the-art analysis).
- You have conducted experiments to show the effectiveness of your approach. Make sure you justify your choice of metrics.
- You have learned some new NLP tools and techniques.

Overly long notebooks with little substance will be penalised. If mature software already exists to solve your problem, it is not sufficient to simply submit this software as the solution. You should try and improve on the solution, or implement the solutions for another language. Black Box solutions are frowned upon — you should be able to explain to your manager how the model works and what its limitations are, in particular what it gets wrong and why. We suggest that you avoid complex neural network approaches for this exercise. Spending all of your time tuning the parameters of a complex model will not be highly evaluated.

8 List of Deadlines

Here is a list of the deadlines and what should be done by each deadline:

21.10.2021, 23:55 — All group members must be registered for their project group in TUWEL

22.10.2021, 23:55 — Select the topic from Section 5 that your group will work on and fill out the Google form via the link from TUWEL

11.11.2021, 23:55 — Deadline for pushing Milestone 1 to GitHub

02.12.2021, 23:55 — Deadline for pushing Milestone 2 to GitHub

6.1.2022, 23:55 — Deadline for reserving a time slot for the final presentation in TUWEL

13.1.2022, 23:55 — Deadline for pushing your presentation material

14.1.2022 — 9:00-12:00, 13:00-16:00 Presentations from all groups — Online on Zoom. Share your own slides, the uploaded presentation slides can be used as a back-up in case of technical problems.

20.1.2022, 23:55 — Deadline for pushing your final submission to GitHub

9 Office hours

Allan Hanbury Thursdays, 13:00-14:00
(see changes on this TISS page: https://tiss.tuwien.ac.at/person/48222).

Gábor Recski Wednesdays, 15:00-16:00
(see changes on this TISS page: https://tiss.tuwien.ac.at/person/336863).

References

[1] Pablo Accuosto, Mariana Neves, and Horacio Saggion. "Argumentation mining in scientific literature: from computational linguistics to biomedicine". In: *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36.* CEUR Workshop Proceedings. 2021.

[2] Valerio Basile et al. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://aclanthology.org/S19-2007.

[3] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://www.aclweb.org/anthology/D15-1075.

[4] Gjorgjina Cenikj, Tome Eftimov, and Barbara Koroušić Seljak. "SAFFRON: tranSfer leArning For Food-disease RelatiOn extractioN". In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, June 2021, pp. 30–40. DOI: 10.18653/v1/2021.bionlp-1.4. URL: https://aclanthology.org/2021.bionlp-1.4.

[5] Ido Dagan, Oren Glickman, and Bernardo Magnini. "The PASCAL Recognising Textual Entailment Challenge". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Ed. by Joaquin Quiñonero-Candela et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 177–190.

[6] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Version 1. In: *arXiv preprint arXiv:1810.04805* (Oct. 11, 2018). arXiv: 1810.04805v1 [cs.CL]. URL: http://arxiv.org/abs/1810.04805v1.

[7] Anca Dumitrache, Lora Aroyo, and Chris Welty. "Crowdsourcing Ground Truth for Medical Relation Extraction". In: *ACM Transactions on Interactive Intelligent Systems* 8.2 (2018), 1–20. ISSN: 2160-6463. DOI: 10.1145/3152889. URL: http://dx.doi.org/10.1145/3152889.

[8] Tim Graf and Luca Salini. "bertZH at GermEval 2019: Fine-Grained Classification of German Offensive Language using Fine-Tuned BERT". In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019, pp. 434–437.

[9] Yufan Guo et al. "Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review". In: *Bioinformatics* 29.11 (2013), pp. 1440–1447.

[10] Yufan Guo et al. "Identifying the information structure of scientific abstracts: an investigation of three different schemes". In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. 2010, pp. 99–107.

[11] Suchin Gururangan et al. "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://www.aclweb.org/anthology/N18-2017.

[12] Iris Hendrickx et al. "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006.

[13] María Herrero-Zazo et al. "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions". In: *Journal of Biomedical Informatics* 46.5 (2013), pp. 914–920. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.07.011. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001123.

[14] Sarvnaz Karimi et al. "Cadec: A corpus of adverse drug event annotations". In: *Journal of Biomedical Informatics* 55 (2015), pp. 73–81. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2015.03.010. URL: https://www.sciencedirect.com/science/article/pii/S1532046415000532.

[15] Ritesh Kumar et al., eds. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), May 2020. ISBN: 979-10-95546-56-6. URL: https://aclanthology.org/2020.trac-1.0.

[16] Nut Limsopatham and Nigel Collier. "Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1014–1023. DOI: 10.18653/v1/P16-1096. URL: https://aclanthology.org/P16-1096.

[17] Thomas Mandl et al. "Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages". In: *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR, 2021. URL: http://ceur-ws.org/.

[18] Thomas Mandl et al. "Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages". In: *Proceedings of the 11th Forum for Information Retrieval Evaluation*. FIRE '19. Kolkata, India: Association for Computing Machinery, 2019, 14–17. ISBN: 9781450377508. DOI: 10.1145/3368567.3368584. URL: https://doi.org/10.1145/3368567.3368584.

[19] Thomas Mandl et al. "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German". In: *Forum for Information Retrieval Evaluation*. FIRE 2020. Hyderabad, India: Association for Computing Machinery, 2020, 29–32. ISBN: 9781450389785. DOI: 10.1145/3441501.3441517. URL: https://doi.org/10.1145/3441501.3441517.

[20] Marco Marelli et al. *The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment*. May 2014. DOI: 10.5281/zenodo.2787612. URL: https://doi.org/10.5281/zenodo.2787612.

[21] Zulfat Miftahutdinov and Elena Tutubalina. "Deep Neural Models for Medical Concept Normalization in User-Generated Texts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 393–399. DOI: 10.18653/v1/P19-2055. URL: https://aclanthology.org/P19-2055.

[22] Julian Risch et al. "Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments". In: *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*. Düsseldorf, Germany, 2021, pp. 1–12. DOI: 10.48415/2021/fhw5-x128. URL: https://netlibrary.aau.at/urn:nbn:at:at-ubk:3-798.

[23] Julia Struß et al. "Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language". In: *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*. München, Germany: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, Oct. 2019, pp. 352–363. URL: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/GermEvalSharedTask2019Iggsa.pdf.

[24] O. Taboureau et al. "ChemProt: a disease chemical biology database". In: *Nucleic Acids Res* 39.Database issue (2011), pp. D367–372.

[25] Aarne Talman and Stergios Chatzikyriakidis. "Testing the Generalization Power of Neural Network Models across NLI Benchmarks". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 85–94. DOI: 10.18653/v1/W19-4810. URL: https://www.aclweb.org/anthology/W19-4810.

[26] Simone Teufel and Marc Moens. "Summarizing scientific articles: experiments with relevance and rhetorical status". In: *Computational linguistics* 28.4 (2002), pp. 409–445.

[27] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. "Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics". In: *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, pp. 1493–1502.

[28] Simone Teufel et al. "Argumentative zoning: Information extraction from scientific text". PhD thesis. Citeseer, 1999.

[29] Elena Tutubalina et al. "Medical concept normalization in social media posts with recurrent neural networks". In: *Journal of Biomedical Informatics* 84 (2018), pp. 93–102. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.06.006. URL: https://www.sciencedirect.com/science/article/pii/S1532046418301126.

[30] Ö. Uzuner et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text". In: *J Am Med Inform Assoc* 18.5 (2011), pp. 552–556.

[31] Adina Williams, Nikita Nangia, and Samuel Bowman. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://www.aclweb.org/anthology/N18-1101.

[32] Marcos Zampieri et al. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 75–86. DOI: 10.18653/v1/S19-2010. URL: https://aclanthology.org/S19-2010.

[33] Marcos Zampieri et al. "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188.

[34] Yuhao Zhang et al. "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 35–45. DOI: 10.18653/v1/D17-1004. URL: https://aclanthology.org/D17-1004.

[35] Maryam Zolnoori et al. "A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications". In: *Journal of Biomedical Informatics* 90 (2019), p. 103091. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.12.005. URL: https://www.sciencedirect.com/science/article/pii/S1532046419300012.