Philipps Universität Marburg

CAISA Lab
https://caisa-lab.github.io/

Course: Introduction to Natural Language Processing

Prof. Dr. Lucie Flek & Dr. Akbar Karimi

Research Datasets

Dataset: *Intended and Perceived Sarcasm*
Link: paper under submission, contact joanplepi@gmail.com for the dataset
Description: Dataset collected from Twitter. Contains around 33K tweets, where 16K tweets are sarcastic and 17K are non-sarcastic. In addition, there are two types of sarcasm, where 12K are intended and 4K are perceived. The dataset contains extra information regarding: 1. Conversational context (extra tweet types in the same thread as the sarcastic tweet, like cue tweet, elicit tweet, and oblivious tweet. 2. Author history (this part might not be possible to share)
Notes: Several ideas for the project can include: exploration, text features, and conversational context features for the sarcasm detection task, analysis of different types of sarcasm 1. Plepi, Joan, and Lucie Flek. "Perceived and intended sarcasm detection with graph attention networks." (previous work in a similar domain/dataset) 2. Current paper regarding the data collection and initial experiments. Draft available upon request
TEAMS:

Dataset: *Misinformation Spreading in Social Media*
Link: https://drive.google.com/drive/folders/1MB6zsrhNerZQlLFBdjJ8sDbvXa2NcELZ
Description: A dataset collected from Reddit, regarding misinformation spreading. It contains a rich amount of information, at both the post level, and aggregated on

the user level. Labels are in several domains like factual information of a post, political bias, science belief, and satire degree.
Notes: As the dataset contains fine-grained scores about both users or posts, there are several ideas that you can explore on different tasks. 1. Plepi. J, Sakketou. F et.al, “FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias” (Introduces the dataset) 2. Plepi. J, Sakketou. F et.al, “Temporal Graph Analysis of Misinformation Spreaders in Social Media”
TEAMS:

Dataset: *Users' Perception*
Link: https://drive.google.com/drive/folders/18iGMBEsQYw8dya9baqhrquQHmpmk71ka
Description: A dataset from r/AITA subreddit on Reddit. The dataset is focused on the moral judgment of users in social networks.
Notes: 1. Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, Yejin Choi “Social Chemistry 101: Learning to Reason about Social and Moral Norms” 2. Ion Stagkos Efstathiadis, Guilherme Paulino-Passos, Francesca Toni, “Explainable Patterns for Distinction and Prediction of Moral Judgement on Reddit” 3. Charles Welch, Joan Plepi, Béla Neuendorf, Lucie Flek, “Understanding Interpersonal Conflict Types and their Impact on Perception Classification”
TEAMS: 12

Dataset: *Medical Named Entity Recognition*
Link: https://drive.google.com/drive/folders/1-iVd2E3DQNq3pIWCUrVP1P3z37YCeFTI
Description: This task will focus on the recognition of disease mentions in tweets written in Spanish after selecting primarily first-hand experience of diseases and other health-relevant content (from patient associations and professional healthcare institutions). The aim is to use social media as a proxy to better understand the societal perception of disease, from rare immunological and genetic diseases such as

cystic fibrosis, highly prevalent conditions such as cancer and diabetes, to often controversial diagnoses such as fibromyalgia and even mental health disorders.

Automatic data selection actively retrieved posts with personal messages and from patient associations. Thus, the SocialDisNER shared task will enable training deep learning named entity recognition approaches to detect all kinds of disease mentions in social media, including both lay and professional language.

Notes: If you use any data from this repository, please cite our scientific paper instead of the Zenodo repo:

Luis Gasco Sánchez, Darryl Estrada Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. The SocialDisNER shared task on the detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines, and corpora.

In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 182–189, Gyeongju, Republic of Korea. Association for Computational Linguistics.

TEAMS:

Dataset:

Link: https://github.com/akkarimi/BERT-For-ABSA/tree/master/ae

Description: Given a collection of review sentences, the goal is to extract all the terms, such as “waiter”, “food”, and “price” in the case of restaurants, which point to aspects of a larger entity [32]. In order to perform this task, it is usually modeled as a sequence labeling task, where each word of the input is labeled as one of the three letters in {B, I, O}. Label ‘B’ stands for “Beginning” of the aspect terms, ‘I’ for “Inside” (aspect terms’ continuation), and ‘O’ for “Outside” or non-aspect terms.
Aspect Extraction Datasets
The dataset includes laptop and restaurant reviews.

Notes: Cite the following if you use this dataset: [1] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq et al., “Semeval-2016 task 5: Aspect based sentiment analysis,” in Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), 2016, pp. 19–30. [2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “SemEval-2014 task 4: Aspect based sentiment analysis,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004
TEAMS:

Dataset: *Aspect Sentiment Classification Dataset*
Link: https://github.com/akkarimi/BERT-For-ABSA/tree/master/asc
Description: Given the aspects with the review sentence, the aim in ASC is to classify the sentiment towards each aspect as Positive, Negative, Neutral. The dataset includes laptop and restaurant reviews.WW
Notes: Cite the following if you use this dataset: [1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “SemEval-2014 task 4: Aspect based sentiment analysis,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004
TEAMS: #6

Dataset: *SemEval 2023 Task 8: Identification of Causal Claims*
Link: https://drive.google.com/drive/folders/1-iVd2E3DQNq3pIWCUrVP1P3z37YCeFTI
Description: **Subtask 1**: Causal claim identification:

For the provided snippet of text, the first subtask aims to identify the span of text that is either a claim, experience, experience_based_claim or a question. These four categories can be defined as follow:
Claim: Communicates a causal interaction between an intervention and an outcome.
Experience: Relates a specific outcome/symptom to an intervention or population based on someone's experience.
Experience-based claim: A claim based on someone's experience.
Question: Poses a question.

Subtask 2: PIO frame extraction:
In this subtask, for a given multi (or single) sentence text snippet and identified claim in that snippet, the task is to extract related Population (P), Intervention (I), and Outcome (O) frames. While it is rare, it may be the case that there is more than one claim in any given post. In that case, we want to identify PIO elements for a given claim. This can be framed as a sequence tagging task.

Notes: Cite the following if you use this dataset:
[1] Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. 2023. Semeval-2023 task 8: Causal medical claim identification and related PIO frame extraction from social media posts. In Proceedings of the 17th International Workshop on Semantic Evaluation, Toronto, Canada. Association for Computational Linguistics.
[2] Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. Redhot: A corpus of annotated medical questions, experiences, and claims on social media. In European Association of Computational Linguistics (EACL).

TEAMS:

Dataset: *(CLEF) Many sets of challenges*

Link: https://clef2022.clef-initiative.eu/index.php?page=Pages/labs.html

Description: CLEF promotes the systematic evaluation of information access systems, primarily through experimentation on shared tasks. CLEF 2022 consists of a set of 14 Labs designed to test different aspects of multilingual and multimedia IR systems.

Notes: *There are many different tasks and datasets in this set that are not possible to upload to Google Drive*.

TEAMS:

Team #7: ARQMath: Answer Retrieval for Questions on Math

Dataset: *Empathy in Text-based Mental Health Support*
Link: https://arxiv.org/pdf/2105.06829.pdf https://github.com/behavioral-data/Empathy-Mental-Health
Description: A dataset of peer support interactions on Reddit, annotated with empathy labels and ratings.
Notes: Label types: Emotional reactions, explorations, interpretations
TEAMS: 14

Dataset: *EmpatheticDialogues*
Link: https://github.com/facebookresearch/EmpatheticDialogues
Description: Conversations in which participants are discussing situations in which they felt specific emotions. This is a heuristic approach for collecting “empathic” interactions.
Notes: Paper: https://arxiv.org/abs/1811.00207
TEAMS:

Dataset: *WASSA 2022/2023 Datasets*
Link: https://codalab.lisn.upsaclay.fr/competitions/11167#learn_the_details-datasets
Description: Dataset of empathetic news reactions. Participants read a news article about something traumatic experienced by others and write a response. The labels are empathy and distress. There are also other labels for emotions and personality indicators. The 2023 WASSA task released conversations between two participants who read the same article.
Notes: You may need to register for the task in order to have access to the datasets, but they might be available without registration for the 2022 competition: https://codalab.lisn.upsaclay.fr/competitions/834#learn_the_details-datasets
TEMAS: 2

Dataset: *SPINOS: A Dataset of Subtle Polarity and INtensity Opinion Shifts*

Link: https://github.com/caisa-lab/SPINOS-dataset
Description: A stance detection dataset on sociopolitical topics discussed on Reddit.
Notes: The dataset is introduced and analyzed in their paper: Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion, Flora Sakketou, Allison Lahnala, Liane Vogel and Lucie Flek
TEAMS:

Dataset: *An Expert-Annotated NLP Dataset for Legal Contract Review (CUAD)*
Link: https://www.atticusprojectai.org/cuad
Description: Many specialized domains remain untouched by deep learning, as large labeled datasets require expensive expert annotators. They address this bottleneck within the legal domain by introducing the Contract Understanding Atticus Dataset (CUAD), a new dataset for legal contract review. CUAD was created with dozens of legal experts from The Atticus Project and consists of over 13,000 annotations. The task is to highlight salient portions of a contract that are important for a human to review. We find that Transformer models have nascent performance, but that this performance is strongly influenced by model design and training dataset size. Despite these promising results, there is still substantial room for improvement. As one of the only large, specialized NLP benchmarks annotated by experts, CUAD can serve as a challenging research benchmark for the broader NLP community.
Notes: The dataset is introduced and analyzed in their paper.
TEAMS:

Dataset: *AmbiEnt Dataset*
Link: https://github.com/alisawuffles/ambient
Description: A linguist-annotated dataset of ambiguous language.
Notes: The dataset is introduced and analyzed in their paper.
TEAMS:

Dataset: *FEVER*
Link: https://fever.ai/dataset/fever.html
Description: FEVER (Fact Extraction and VERification) consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as **Supported**, **Refuted**, or **NotEnoughInfo**. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment.
Notes: FEVER: a large-scale dataset for Fact Extraction and VERification James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal
TEAMS: #10

Dataset: *CheckThat Lab - Check-worthiness Detection*
Link: https://gitlab.com/checkthat_lab
Description: The task is to identify check-worthy social media posts that contain likely misinformation, evoke polarization, and hate speech. The datasets (CheckThat 2021 and 2021) are multilingual and mostly about Covid-19
Notes: The students can do research on how they can learn multilingual models for zero-shot learning languages in a cost-efficient way. For instance investigating cultural, and geopolitical similarities of the countries for selecting as a training set, injecting background information (e.g political ontology) about the country into transformers for performance increase. Link to the related papers: 1- Schlicht, I.B., Flek, L., Rosso, P. (2023). Multilingual Detection of Check-Worthy Claims Using World Languages and Adapter Fusion https://doi.org/10.1007/978-3-031-28244-7_8 2- Nakov, P. *et al.* (2022). Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection https://doi.org/10.1007/978-3-031-13643-6_29 3- Nakov, P. *et al.* (2021). Overview of the CLEF–2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. https://doi.org/10.1007/978-3-030-85251-1_19
TEAMS:

Dataset: *CheckThat Lab - Subjectivity Detection*
Link: https://gitlab.com/checkthat_lab - 2023 edition task 2
Description: The task is to detect whether a news article's sentence is subjective or objective. As a check-worthy detection task, this task is also multilingual.
Notes: The same research questions as the check-worthy detection task could be investigated. - The dataset paper: https://dl.acm.org/doi/abs/10.1007/978-3-031-28241-6_59
TEAMS:

Dataset: *NLP Methods for Indoctrination Detection in German History Textbooks*
Link: https://studtudarmstadtde-my.sharepoint.com/:x:/g/personal/larsmatthias_wolf_stud_tu-darmstadt_de/EQfzGN6XHrpPkD4ZBwbZbXABsgRiMXJlVFtQoTtEHmkIyQ?rtime=PZOPappE20g https://history-analysis.herokuapp.com/embeddings
Description: Controlling information and mass media is crucial for dictators to stay in power. While propaganda and fake news detection have seen a surge in research attention lately, this work focuses on analyzing deeper beliefs and values. As a collaboration between political science and computer science, we introduce the novel task of indoctrination detection. We processed 46 scanned textbooks from the German Democratic Republic (GDR) and the Federal Republic of Germany (FRG), used in history classes from 1948 to 1989 and covering the two countries' common history from 1900 to after World War II. We automatically analyze these textbooks regarding several facets of indoctrination, which include gatekeeping, selective attribution, subjective language, and appropriation. For examining these, we use embedding-, semantic role labeling- and emotion-based techniques to identify word meaning shifts, activity, and passivity of entities and emotions towards entities in the textbooks. We then create a corpus for the new task of indoctrination detection by manually annotating 336 excerpts of the history textbooks for indoctrination mechanisms and entities affected.
Notes: Lucie Flek (Universität Marburg, DE) Ivan Habernal (TU Darmstadt, DE) Christopher Klamm (Universität Mannheim, DE) Dani Sandu (European University Institute, IT) Lars Wolf (TU Darmstadt, DE)
TEAMS:

Dataset: *Formality style transfer (GYAFC)*
Link: https://github.com/raosudha89/GYAFC-corpus
Description: Style transfer is the task of automatically transforming a piece of text in one particular style into another. A major barrier to progress in this field has been a lack of training and evaluation datasets, as well as benchmarks and automatic metrics. In this work, we create the largest corpus for a particular stylistic transfer (formality) and show that techniques from the machine translation community can serve as strong baselines for future work. We also discuss challenges of using automatic metrics.
Notes: See the paper: https://arxiv.org/abs/1803.06535 You have to request it by following the instructions on GitHub
TEAMS:

Dataset: *Stanford Politeness*
Link: https://convokit.cornell.edu/documentation/stack_politeness.html
Description: They propose a computational framework for identifying linguistic aspects of politeness. Their starting point is a new corpus of requests annotated for politeness, which they use to evaluate aspects of politeness theory and to uncover new interactions between politeness markers and context.
Notes: See publication here: https://nlp.stanford.edu/pubs/politeness.pdf
TEAMS:

Dataset: *Rude and Supportive Comments*
Link: https://github.com/hadarishav/Ruddit
Description: On social media platforms, hateful and offensive language negatively impact the mental well-being of users and the participation of people from diverse backgrounds. Automatic methods to detect offensive language have largely relied on datasets with categorical labels. However, comments can vary in their degree of offensiveness. They create the first dataset of English language Reddit comments that has fine-grained, real-valued scores between -1 (maximally supportive) and 1 (maximally offensive). The dataset was annotated using Best--Worst Scaling, a form of comparative annotation that has been shown to alleviate known biases of using rating scales.

Notes: See the paper: https://arxiv.org/abs/2106.05664
TEAMS: #11

Dataset: *RealToxicityPrompts and Toxicity Data*
Link: https://allenai.org/data/real-toxicity-prompts
Description: They investigate the extent to which pre-trained LMs can be prompted to generate toxic language and the effectiveness of controllable text generation algorithms at preventing such toxic degeneration. They create and release RealToxicityPrompts, a dataset of 100K naturally occurring, sentence-level prompts derived from a large corpus of English web text, paired with toxicity scores from a widely-used toxicity classifier.
Notes: The paper link has code and data. Paper: https://aclanthology.org/2020.findings-emnlp.301
TEAMS:

Dataset: *Conversational Receptiveness*
Link: https://osf.io/2n59b/?view_only=
Description: How can we be more receptive when we talk to people we disagree with?
Notes: See the paper: https://www.mikeyeomans.info/papers/receptiveness.pdf
TEAMS: