1. Data Set:
   1. Articles and Posts in German
   2. Posts are (partially) labeled
   3. Articles and Posts give rise to a tree structure, as each post replies to either an article or a post
   4. Posts and articles are completely saved.
   5. As our documents are user comments (mostly), they may contain things such as:
      1. Typos
      2. Auto-correct fails (or, maybe just somewhat incoherent speech in general)
      3. Creative ways of putting emphasis, e.g. Capslock or repeated vocals (looooooooooooooong)
      4. Sidestepping Curse word filters by replacing letters, for example by stars.
      5. smileys (no exotic non-ASCII symbols)
      6. No HTML-Tags or more complicated structure, though maybe links.
   6. Articles are more difficult, having a HTML-document structure.
   7. 9 Labels:
      1. Arguments used
      2. Personal stories
      3. Discriminating
      4. Inappropriate
      5. Offtopic
      6. Sentiment (negative/neutral/positive)
      8. Some of the labels are very unbalanced
         1. Example:  
            Inappropriate (~3000 no-Instances, ~300 yes-Instances)  
            SentimentPositive (~3500 no-Instances, ~40 yes-Instances)
            1. SentimentPositive has so few positive samples, it’s basically untrainable.
   8. 3600 of the posts have all 9 labels set.   
      Most have only a subset of the labels set. Those with some labels set make up ~40,000 posts.   
      The remaining ~1 million posts are unlabeled.
   9. With Exception of the post-sentiment, each sample can have any combination of labels.  
      (and each post has at most one of the labels SentimentPositive, -Neutral, -Negative set to 1)
   10. There are some additional labels, which are given by the forum, most notably:
       1. Positive & negative votes
       2. Whether a post was deleted
2. Short statistic analysis of the features:
   1. Correlation matrix (between those samples which had all 9 labels set)  
      [Kovarianz-Matrix-zeigen]
      1. The correlation between most features is negligible;   
         Only exception: SentimentNegative and SentimentNeutral, which has negative correlation (though smaller than excepted)  
         Je nach Korrelation: Negativ klar, da mehr oder weniger disjunkt (aber nicht annähernd so ausgeprägt wie sein sollte, nämlich 1 für disjunkt!)  
         Positiv: Dafuq.  
         In jedem Fall ist diese Korrelation die Erklärung, warum wir alle Labels als Binärlabels ansehen, und keinen Regressor für Sentiment bauen.
3. Used features:
   1. Word embeddings, for each sample (i.e. labeled post) one: (… we tried both i. and ii.)
      1. Word-Embedding as given by Spacy (dense embedding) (we tried both md and lg [latter has many more word-embeddings], but lg was in fact not consistently better in results)
      2. Word-Embedding yielded by calculating the tf-idf, after lemmatizing and removing punctuation (sparse embedding; lemmatization ‘necessary’ because of way too few samples for the many forms of German words)
4. Trained classifiers:
   1. SVC with different kernels, most notably the RBF-Kernel and polynomial kernels of different degrees. (was is RBF???)
   2. Boosted Decision Trees
   3. MLP (Feed-Forward NN)
   4. k-Nearest Neighbor
   5. Naive Gaussian Bayes
   6. Logistic Regression
5. Results:
   1. [Show graphs of best models; With meta-labels and without]
      1. Especially with sparse high dimensional embeddings, primal/dual SVC did give best results
      2. With dense embeddings, SVC with RBF-Kernel and other classifiers gave best results (some that, some that)
   2. Using the meta-labels Positive-/Negative-votes and whether post was deleted, didn’t improve the results
   3. Difficult labels and somewhat simple labels:
      1. Difficult: Discriminating, Inappropriate, Offtopic
         1. The classifiers for these all might benefit from knowing the full parent comment thread as well as the parent post.
         2. Especially offtopic intuitively should need the original post to be labeled – however it wasn’t the one the classifiers struggled with the most, which was (?) Discriminating(?)
            1. However, there isn’t any significant improvement when using
         3. Unsure how to use the tree structure to help labels ‘Discriminating’ and ‘Inappropriate’.
         4. For label ‘Offtopic’, we can give the classifier both the embedding for the post as well as the embedding of its parent article.
            1. We did this by tidying up the article (removing HTML), turning it into a series of paragraphs, then calculated for each paragraph the similarity to the post in question, and then gave the classifier the highest similarity we obtained in this process.  
                 
               However, this procedure gave no significant improvement.
   4. Are our results good/bad?
      1. Objectively, the classifiers for most labels aren’t really useful.  
         PersonalStories and PossiblyFeedback are probably reliable enough to be used in non-critical applications.
      2. Comparatively: We aren’t the first to evaluate the dataset.  
         [Show Other-Authors-file]  
         Their results are from 2017, and even though they used Long-Short-Term-Memory and alternative Word-Embeddings (giving Embeddings to documents as whole), their results are on par with ours, with ours being in general a little higher   
         (No matter whether we look at only the text as input, or text + votes + post-status)