Task 3: Sentiment analysis

Data

1. A subset of *Yelp reviews* from this study:
   Zhang, Xiang, Junbo Zhao, and Yann LeCun. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems* 28.

2. *SentiWords* lexicon:
   Gatti, L., Guerini, M., & Turchi, M. (2015). SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, *7*(4), 409-421.

Subtasks and points

1. Familiarize yourself with the Yelp data. Split the provided train set into (new) train (70% of the *whole* dataset) and validation (10% of the *whole* dataset) subsets. Describe the data. (5)
   Note: We will reuse the splits in the next labs, so fix the random state.
2. The starter notebook uses *CountVectorizer* to convert texts into vectors and MultinomialNB classifier. Experiment with different vectorizer parameters (use of stopwords, minimum document frequency, binary features, and lowercasing) and the smoothing parameter of the NB classifier. Select the best configuration based on the validation set, apply it to the test set. (25)
   a. Experiment with a different classifier (e.g. SVM). (10)
3. Familiarize yourself with the *SentiWord* lexicon. Process and describe the data. (5)
4. Develop a lexicon-based sentiment classifier using *Stanza* for lemmatization and POS-tagging. (Mind difference in labeling: sentences: 0 – negative, 1 – positive; words: continuous scores from the range [-1, 1]. Note that *SentiWords* and *Stanza* use different POS tag sets.) (35)
   a. Use validation set to optimize the threshold value for binary classification. (10)
5. Summarize and compare the evaluation results (accuracy on the test set) of all tested configurations. Analyze misclassified examples. (10)