NLP CHATBOT REPORT

by Sanne Hoeken

October 25, 2020

Introduction to Human Language Technology

Student number: 2710599

1. Introduction

For this project I have set up a framework in which a chatbot can be built that can respond to the emotion and topic of an utterance. In the building process, the chatbot is provided with a classification model that can detect emotions and a classification model that can detect topics. The latter model uses a set of keywords that are linked to a topic. These keywords are included in a responder file. Using the same responder file, the chatbot can return an associated response based on the detected emotion and topic. In the framework that I have set up, different classification models can easily be implemented to see what the consequences are for the performance of the chatbot. The framework (including training data and all built models) can also be consulted as a repository on Github via the following link: https://github.com/SanneHoeken/NLP-Chatbot-Framework

2. Design

Both the emotion and the topic classification models include many parts that offer many variation possibilities. For this project I selected a set of settings and for each setting a set of variation options, which I will explain in the following sections. I set up the framework in such a way that you can easily adjust these settings. In Python I created Classes for both the emotion classifier and the topic classifier. The settings and the associated preprocessing models which I will explain in the next section can be attached as attributes to these Classes. In case of the emotion classifier, the preprocessing models can be created by another Class: an emotion classification model trainer Class (yes, that's a mouthful). The settings can again be attached to this Class as an attribute, just like the training data. The Class's methods can build and store the models needed to classify an emotion in a way that meets all settings. The classifier Classes contain the methods that predict the emotion or the topic of an utterance based on the attached settings and models.

2.1 Emotion Classification Settings

For the emotion classification model, I developed eight different settings. I will now explain the variation options for each.

1. Source (s) of training data. The relevance and quality of training data is crucial for the prediction performance of the model. For this classification task, the Multimodal EmotionLines Dataset (MELD) is used as training data. This dataset consists of approximately 13,000 utterances from 1,433 dialogues from the TV series Friends. Each utterance is annotated with emotion and sentiment labels, and encompasses audio, visual, and textual modalities (Poria et al 2019, 527). The emotion classification model only uses utterances in text form with emotion annotations. The utterances in each dialogue were annotated with Ekman's six universal emotions (Joy, Sadness, Fear, Anger, Surprise, and Disgust) plus a Neutral label. (Poria et al 2019, 529). So, the default setting for the training data is the use of the MELD which seems to be very suitable for the detection of emotions within conversations. With an adjustment of the setting, the training data can be extended with a non-

conversational dataset: the Tweet Emotion Intensity Dataset (which I will refer to as TEID). This dataset consists of four datasets of tweets annotated for intensity of anger, joy, sadness, and fear (Mohammed & Bravo-Marquez 2017). I have only included the annotated tweets whose intensity value was higher than 0.5. This is an arbitrary choice and follow-up research should determine the effects of a certain threshold value. After adjusting this training data setting, the utterances and labels from the chosen dataset(s) (consisting of CSV files) are imported using the *pandas* package, and passed to the model trainer class.

2. Balance of training data. Analysis of the training data shows that the distribution of the training utterances across the emotion classes is skewed. A table and visualisations of this analysis are added as Appendix 1. Neutral is a dominant class and other emotions (e.g. disgust and fear) are minority classes for both the MELD and the MELD extended with the TEID (although to a lesser extent). A consequence of this skewed distribution may be moderate recall for predicting the underrepresented emotions. The second parameter that I therefore implemented concerns the possibility of balancing the training data. To balance, a Python Class is imported from the *imbalanced-learn* toolbox to perform over-sampling using Synthetic Minority Over-Sampling Technique (SMOTE). It aims to balance class distribution by randomly increasing minority class utterances by replicating them. A distribution of the dataset after this resample method can also be seen in Appendix 1.

3. Feature representation. To use a machine learning model, utterances must be transformed into numerical representations. Within the feature representation setting, a choice can be made between three different types of representations: bag-of-words vector representations, TF-IDF vector representations and word embeddings. A.o. modules from the *Scikit-learn* package make these transformations possible. The CountVectorizer module turns a text dataset into a bag-of-words representation. The TfidfTransformer module takes the bag-of-words representations as input and converts them to TF.IDF values. For the transformation to word embeddings, a pre-trained embedding model is loaded online using the *gensim* api (during training of the classification model). The embedding representation is then created for each utterance by extracting the word embedding for each token in the utterance, adding them together and taking the average over it. The preprocessing models used for the vector transformations are saved, so that they can easily be passed to the emotion classifier Class.

4. Word embedding model. As just described, word embedding transformations require a pre-trained embedding model. Within this setting one can choose a pre-trained model that is part of the *gensim-data* project. Since words not in the embedding model's vocabulary are ignored, it matters what type of text the embedding model is trained on: news, Wikipedia, Twitter, spoken dialogues. This can be adjusted within this setting.

5. Vector dimensions. An adaptable part of training an embedding model is the dimensionality of each vector. More dimensions require more data, but can lead to more accurate models. A reasonable number of dimensions are 100 to 500. With the choice of a pre-trained model from the *gensim-data* project, the vector dimensionality is also set.

6. Filtering stopwords. The stopwords in a language do not add much meaning and will therefore probably not be able to provide much information to the classification model. Therefore, when transforming utterances into numerical representations, you can set to filter stopwords. The collection of English stopwords is imported from the *nltk* package.

7. Frequency threshold. In addition to stop words, utterances can also be filtered on the basis of a frequency threshold that indicates how often a word must at least occur in the dataset. Words that occur only once or twice (or perhaps up to tens of times) in a dataset are

unlikely to make a meaningful contribution to the training of the model, so it is probably better to filter them.

8. Classifier. Once a training data selection has been transformed into numerical representations, the task is to build a classifier. Within this classifier setting two different machine learning algorithms can be chosen: Naive Bayes Classification and Linear Support Vector Classification. The differences between these classifiers are beyond the scope of this project and I will therefore not discuss them. Both of the algorithms are implemented using *Sci-kit learn* modules.

2.2 Topic Classification Settings

The topic classification models follow a different structure than the emotion classification models. Instead of transforming text into numerical representations and then having a machine learning algorithm, trained on labeled data, make a prediction, the topic classification models create a set of words similar to the words in the utterance and then try to match the total collection of words with a set of keywords that are linked to a particular topic. In this framework, the topic with the most matched keywords forms the prediction. The mappings from keywords to topics are attached as an attribute to the classifier class. For the topic classification model, I developed nine settings. I will now explain the variation possibilities for each.

1. Semantic model. A semantic model is used to create a set of similar words. Within this setting one can choose a word embedding model, Wordnet, or a combination of these two model types. Every word embedding in an embedding model captures relationships with other words in the corpus on which the model is trained: words with similar distributions have similar vectors. Wordnet, on the other hand, is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Nouns, verbs, and adjectives of a certain language are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. The Wordnet database is imported from the *nltk* package.

2. Word embedding model. 3. Filtering stop words. These two settings include the same possibilities as described in the previous section.

4. Size of the word neighborhood. Within this setting, the amount of similar words obtained for each token can be adjusted. If the similar words are obtained through both an embedding model and Wordnet, half of the neighborhood will be taken from each of the two sources. The greater the number of similar words, the greater the chance of a match with a keyword. However, this may be at the expense of the quality of the match between the keyword and the original token. So, with this setting an attempt can be made to find the right balance.

5. Utterance processing. Obtaining similar words for each word in an utterance requires converting the utterance into a list of word-like units. In this setting, the processing method can be adapted to tokenization or lemmatization. The methods use the word_tokenize function and the WordNetLemmatizer method from the *nltk* package respectively.

6. Wordnet hyponyms. 7. Wordnet hypernyms. When using WordNet to create a set of similar words, the lemmas of all synonym sets of a token are collected. Each synonym set has different relationships with other so called synsets. If one these settings is changed to True, all lemmas of the synsets that are hyponymically or hypernymically related to each

synomym set of the token are also added to the set of similar words. After all similar lemmas have been collected, a similarity measurement is used to determine which lemmas belong to the top N (size of the word neighborhood) similar words.

8. Similarity metric. Matching a word from the collection of similar words with a keyword is done on the basis of a similarity measurement. There are different metrics for this, corresponding to the different types of semantic models. Within this setting, the similarity can be either quantified by taking the vector representations for two words from an embedding model and comparing them through the cosine similarity function. Or, for two words the synsets are collected and each pair is provided with a Wordnet score. Ultimately, the highest score is retained. There are many options for the latter metric. In this setting a choice can be made between three path-based measures: Shortest path, Wu & Palmer and Leakcock & Chodorow. The main idea of path-based measures concerns a function of path length linking the synsets and the position of the synsets in the Wordnet taxonomy (Meng, Huang & Gu 2013).

9. Similarity threshold. The determination of whether the similarity between a word and a keyword is sufficient for a match depends on the similarity threshold that is set. If the result of the similarity measurement is above this threshold, it is decided that there is a match.

2.3. Responder

As discussed briefly in the introduction, the chatbot system uses a responder file to match utterances to keywords for the prediction of a topic, and to return a response based on the predicted topic and emotion. The structure of my responder file is as follows: the object 'topics' contains an array of objects that each contain a topic value, an array of keywords, and an array of emotions. Each emotion is an object that contains an emotion value and an array of responses. In addition to the given topics animals, people, sports, food, & places, I have added an extra topic 'other' that is predicted if there is no matching with a keyword. For each of the topics (except 'other') I made a list of ten keywords. To make this list I used the following semantic model: my personal intuition. I will come back to this in the discussion.

3. Models & Hypotheses

If I wanted to test the effects of all the different variation possibilities, I end up with hundreds of models for both classification tasks. And that leaves me with the ability to test ten to hundreds of thousands of different chatbot systems. Although the architecture of this framework does not make this unfeasible, it seems to me a bit outside the scope of this project to test all possible variation. Nevertheless, I have implemented all these settings anyway because it seems very useful to me to have different text classification models within reach in an NLP research career. For this project I limited myself to five variations, and thus five models, per classification task. Schematic overviews of the settings of the models are added as Appendix 2, but I will also describe them in the following sections.

3.1 Emotion classification models and hypotheses

I trained the first model on the unbalanced MELD. The utterances are represented as embeddings using word embeddings created in the GloVe project trained on the Wikipedia 2014 and Gigaword 5 corpora, and consist of 300 dimensions. Stopwords are ignored and the

frequency threshold is set to 4. A linear SVM classifier is trained to make the predictions. In the second model, the training data is balanced as described in the settings. All other settings remain the same. I expect (as also mentioned in the settings) that the effect of this can mainly be seen in an improved recall for the emotions other than neutral, and in particular minority emotions such as fear, disgust and sadness. In addition, I think that the precision of the dominant neutral emotion is improved because fewer utterances are incorrectly annotated as neutral. However, the recall will likely be reduced.

In the third and fourth model, different word embedding models are used. In the third model word embeddings are used that are trained on the Google News corpus with still a dimensionality of 300. Where the previous model is trained on both Wikipedia articles and written news, this model is only trained on the latter genre. The genre diversity is less but the size of the vocabulary is greater: the Google News model contains almost ten times more vectors. I think that the language used in written news is more related to spoken utterances (as in the MELD) than Wikipedia articles and therefore the reduced genre diversity makes little difference. I think the larger vocabulary has positive effects on the overall accuracy of the model, as it probably means that more data can be used during training. I expect this effect all the more with the fourth model where pre-trained word embeddings trained on twitter data are used. I think this vocabulary is even more adapted to the MELD data. This model includes 200-dimensional vectors. The other settings for the third and fourth model will be the same as the second model.

In the fifth model, all settings will also be the same as the second model, only with extended training data. As mentioned earlier, the TEID with which the data is extended is not conversational and only contains twitter posts labeled with the emotions anger, joy, sadness and fear. The architecture of the chatbot is built in such a way that each utterance is classified separately. I think tweets are more comparable with the input the chatbot eventually gets: utterances without context. So, in addition to the fact that there is more data on which the model is trained, I think the extended data collection also consists of more relevant data, which will both benefit the accuracy of the classification model.

3.2 Topic classification models and hypotheses

In the first topic classification model I built, a set of similar words is created via an embedding model with 300-dimensional word embeddings trained on Wikipedia 2014 and Gigaword 5 data. Stopwords are ignored, the word neighborhood is 10, and utterances are processed by tokenization. The similarity metric is the cosine-similarity function and the threshold is set to 1.0 (so we could roughly speak of exact matching). The second model obtains similar words with Wordnet instead of embeddings. Hyponymic and hypernymic relationships between the synsets are also included. The other settings remain the same as the first model. It is difficult to make a prediction about the consequences of this variation because it is not a foregone conclusion whether similarity based on human intuition (Wordnet) is better than similarity based on sampling of empirical data (word embeddings) or vice versa. I think a combination of both would be best. And that is how the similar words in the third model are obtained, both via Wordnet and a word embedding model. Again, hyponymic and hypernymic are included and all other settings are the same as the previous models.

Fares et al. (2017) proved in their research that lemmatization of text improves the performance of a model that performs a semantic similarity task by means of word embeddings. Therefore, I think this improvement will also apply to my fourth model in which

I adjusted the processing setting from tokenization to lemmatization and I kept the other settings the same as the third model. In the last model, the settings of model four have been adopted, but the similarity metric has been adjusted to the shortest path-based measure from Wordnet. I think that the same argument can be cited here as with the difference between the first and second models, and that it is therefore difficult to predict the effect of this setting adjustment.

4. Results & Conclusions

To determine whether the built models are valuable for performing a classification task at all, I compare the models with a baseline. In case of the emotion classification task, I took the emotion that appeared most common in the data as the result for all predictions as the baseline. For the data within this project, this is the neutral emotion. For the topic classification, I took one of the most frequent topics, people, as a result for all predictions as the baseline. All models have been tested on a provided test set with test utterances.

4.1 Emotion classification reports

The tables below show the performance for the emotion classification models on the test set in the form of precision, recall and F1 measurements for each emotion and averaged over all emotions. Marked in green are the best scores of all models.

Precision	**BL**	**1**	**2**	**3**	**4**	**5**	**Sup.**
Anger	0,00	0,00	0,25	0,33	0,40	0,40	4
Disgust	0,00	0,00	0,25	0,17	0,09	0,25	4
Fear	0,00	0,00	0,25	0,14	0,50	0,38	6
Joy	0,00	0,67	0,75	1,00	0,50	0,50	6
Neutral	0,22	0,25	0,50	0,50	0,40	0,43	8
Sadness	0,00	0,00	0,30	0,17	0,17	0,33	4
Surprise	0,00	0,00	0,00	0,00	0,00	1,00	4
Accuracy	0,22	0,28	0,36	0,25	0,28	0,39	0
Macro avg	0,03	0,13	0,33	0,33	0,29	0,47	36
Weighted avg	0,05	0,17	0,37	0,38	0,33	0,46	36

Table 1: Precision measures for five emotion classification models and baseline (BL)

Recall	**BL**	**1**	**2**	**3**	**4**	**5**	**Sup.**
Anger	0,00	0,00	0,25	0,25	0,50	0,50	4
Disgust	0,00	0,00	0,50	0,50	0,25	0,50	4
Fear	0,00	0,00	0,17	0,17	0,17	0,50	6
Joy	0,00	0,33	0,50	0,33	0,50	0,33	6
Neutral	1,00	1,00	0,38	0,25	0,25	0,38	8
Sadness	0,00	0,00	0,75	0,25	0,25	0,25	4
Surprise	0,00	0,00	0,00	0,00	0,00	0,25	4
Accuracy	0,22	0,28	0,36	0,25	0,28	0,39	0
Macro avg	0,14	0,19	0,36	0,25	0,27	0,39	36
Weighted avg	0,22	0,28	0,36	0,25	0,28	0,39	36

Table 2: Recall measures for five emotion classification models and baseline (BL)

F1	BL	1	2	3	4	5	Sup.
Anger	0,00	0,00	0,25	0,29	0,44	0,44	4
Disgust	0,00	0,00	0,33	0,25	0,13	0,33	4
Fear	0,00	0,00	0,20	0,15	0,25	0,43	6
Joy	0,00	0,44	0,60	0,50	0,50	0,40	6
Neutral	0,36	0,40	0,43	0,33	0,31	0,40	8
Sadness	0,00	0,00	0,43	0,20	0,20	0,29	4
Surprise	0,00	0,00	0,00	0,00	0,00	0,40	4
Accuracy	0,22	0,28	0,36	0,25	0,28	0,39	0
Macro avg	0,05	0,12	0,32	0,25	0,26	0,38	36
Weighted avg	0,08	0,16	0,34	0,26	0,28	0,39	36

Table 3: F1 measures for five emotion classification models and baseline (BL)

First of all, all models show improved scores compared to the baseline results, so the built models seem to be valuable for predicting emotions. Then, coming back to the hypotheses, the results of the first two models show, as predicted, that balanced data has a positive effect on the recall of the emotions other than neutral, and on the precision of the neutral emotion (and actually all emotions), but at the expense of the recall of the neutral emotion. Only the emotion surprise is the exception to this, the prediction performance for this emotion remains just as poor.

The performances of models 3 and 4 compared to the second model seem to indicate that the use of word embedding models trained on Google News or Twitter data does not improve performance. Some scores for some emotions are better but in general the accuracy of the model using the embedding model trained on Wikipedia and Gigaword is the best of all three. This is against all odds. In fact, model 4 was expected to be the most accurate of all three, but this model appears to be the least accurate.

Finally, extending the training data with the TEID seems to provide the most accurate model of all. With the exception of the emotions joy, sadness and neutral, this model scored the highest for precision, recall and F1. An overall performance improvement with the adjustment of this setting confirms the hypothesis. It is also remarkable that this model is the only model that has performance scores higher than zero for the emotion surprise (what a surprise...).

4.2 Topic classification reports

The tables below show the performance for the topic classification models on the test set again in the form of precision, recall and F1 measurements for each topic and averaged over all topic. Again, all models show better results from baseline. So, the built classification models seem to be valuable for predicting topics. In addition, I noticed that the precision for all models is substantially higher than the recall. Nevertheless, almost all scores are higher than for the emotion classification task. Whether the emotion classification task turns out to be more difficult for machines is an interesting question, but due to the large architectural differences between the models, also an issue that will remain unanswered in this project. The results also show that for almost all models the performance on the topics sports and people is much better than on the topics animals and food. Follow-up research should determine whether this effect can be attributed to the settings of the models or to the chosen keywords for these topics.

Precision	BL	1	2	3	4	5	Sup.
Animals	0,00	1,00	1,00	1,00	1,00	0,50	8
Food	0,00	0,33	0,00	0,50	0,50	1,00	7
People	0,22	0,25	0,60	0,50	0,57	0,60	8
Places	0,00	0,80	0,56	0,63	0,63	0,33	8
Sports	0,00	1,00	0,60	1,00	1,00	0,19	5
Accuracy	0,22	0,31	0,33	0,36	0,42	0,31	0
Macro avg	0,04	0,56	0,46	0,60	0,62	0,44	36
Weighted avg	0,05	0,66	0,56	0,71	0,72	0,54	36

Table 4: Precision measures for five topic classification models and baseline (BL)

Recall	BL	1	2	3	4	5	Sup.
Animals	0,00	0,13	0,13	0,13	0,25	0,13	8
Food	0,00	0,14	0,00	0,14	0,14	0,14	7
People	1,00	0,13	0,38	0,38	0,50	0,38	8
Places	0,00	0,50	0,63	0,63	0,63	0,38	8
Sports	0,00	0,80	0,60	0,60	0,60	0,60	5
Accuracy	0,22	0,31	0,33	0,36	0,42	0,31	0
Macro avg	0,20	0,28	0,29	0,31	0,35	0,27	36
Weighted avg	0,22	0,31	0,33	0,36	0,42	0,31	36

Table 5: Recall measures for five topic classification models and baseline (BL)

F1	BL	1	2	3	4	5	Sup.
Animals	0,00	0,22	0,22	0,22	0,40	0,20	8
Food	0,00	0,20	0,00	0,22	0,22	0,25	7
People	0,36	0,17	0,46	0,43	0,53	0,46	8
Places	0,00	0,62	0,59	0,63	0,63	0,35	8
Sports	0,00	0,89	0,60	0,75	0,75	0,29	5
Accuracy	0,22	0,31	0,33	0,36	0,42	0,31	0
Macro avg	0,07	0,35	0,31	0,37	0,42	0,26	36
Weighted avg	0,08	0,39	0,37	0,43	0,49	0,31	36

Table 6: F1 measures for five topic classification models and baseline (BL)

If we compare the first two models, it can be seen that the accuracy score and the average recall scores of the second model are higher (although not much) but the average precision and F1 scores of the first model are higher. If we look at the individual differences (for recall, precision and F1), the second model appears to perform better on the topic people, and the first model appears to perform better on the topics food, places and sports. I think that is an interesting result regarding the difference between Wordnet and an embedding model. Would that mean that human intuition (the source of Wordnet), compared to empirical data (the source word embeddings), is better able to reveal semantic relationships that are about themselves (people) than about other topics? In any case, a combination of these two sources seems to result in a classifier that, as expected, includes the best of both: the overall performance for the precision, recall and F1 is better for the third model compared to the first two models.

The results of the fourth model show that lemmatization of the utterances has a positive effect on the performance of the model. For almost all scores, this model scores substantially higher than the previous models. This confirms the findings of Fares et al. (2017).

Finally, comparison of model 5 with model 4 shows that matching keywords based on a path-based similarity measure in Wordnet, instead of the cosine similarity function, does not lead to any improvement in the model. With the exception of the food topic, no score is better for the fifth model.

4.3 Chatbot test results

For the chatbot systems, I took for each classification task a model that showed the best performance on the test set and a model that differs substantially from this model in terms of settings, but at the same time showed reasonable performance. That leaves me with four different chatbot systems to be compared. The four setups are as follows:

Setup 1: emotion model 5 + topic model 1
Setup 2: emotion model 5 + topic model 4
Setup 3: emotion model 4 + topic model 1
Setup 4: emotion model 4 + topic model 4

Each setup has been tested on all test utterances. For each utterance, the emotion and topic were predicted using the specified models and a response was generated based on the responder file. The results are attached as Appendix 3. A summary of these results can be seen in the table below. The results of the setups are in line with the performances of the classifiers. Setup 2, with the best models for both classification tasks, makes the most correct predictions.

N	**Setup 1**	**Setup 2**	**Setup 3**	**Setup 4**
Correct topic + correct emotion	3	6	2	3
Correct emotion	13	13	10	10
Correct topic	11	15	11	15

Table 7: Test results of four chatbot setups

6. Discussion

As became clear earlier, this framework still leaves a lot of testing possibilities. It would be interesting to try these possibilities in follow-up research so that the most optimal chatbot system with this architecture can be created. In addition, there are many other variation options within the framework that could still be implemented. When building the models, it turned out that combining variation options (e.g. different training data sets and semantic models) can be fruitful. It therefore seems interesting to me to implement the possibility of other combinations, such as a combination of different similarity metrics, a combination of different embedding models and a combination of different synset relationships (e.g. by extension with meronymy and antonomy).

Another part that leaves room for improvement is the selection of keywords for matching topics. The choice of keywords in this project is based solely on the personal intuition of a single individual, myself. Using semantic models instead to find keywords that are semantically related to a topic may increase the chances of matching (a process based on those same semantic models), which may improve the topic classifier's performance. If all of these potential improvements were made, I would test the newly built models on a larger test set (with less spelling errors), that also showed a favorable inter-annotator agreement. I

sometimes got along better with my chatbot system than with the annotators of the current test set.

Finally, expanding the responses in the responder file could lead to a chatbot system that appears more original and creative, in addition to emotional and meaningful. All in all, in this project I have developed a framework that is able to implement different chatbot systems and I have shown some effects of adjusting various settings in such a system.

References

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. “Word vectors, reuse, and replicability: Towards a community repository of large-text resources”. *Proceedings of the 21st Nordic Conference on Computational Linguistics*: 271-276. https://www.aclweb.org/anthology/W17-0237.pdf

Meng, Lingling, Runqing Huang, and Junzhong Gu. 2013. "A review of semantic similarity measures in wordnet." *International Journal of Hybrid Information Technology* 6, no. 1: 1-12.

Mohammad, Saif M., and Felipe Bravo-Marquez. 2017. “Emotion Intensities in Tweets.” *Proceedings of the sixth joint conference on lexical and computational semantics*. http://saifmohammad.com/WebDocs/TweetEmotionIntensities-starsem2017.pdf

Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations”. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*: 527-536. https://www.aclweb.org/anthology/P19-1050.pdf

Appendix 1. Training data analysis

Emotion	MELD	MELD+TEID	Over-Sampled
neutral	4710	4710	4710
joy	1743	2120	4710
surprise	1205	1205	4710
anger	1109	1483	4710
sadness	683	1050	4710
disgust	271	271	4710
fear	268	796	4710
total	*9989*	*11635*	*32970*

Table 8: Frequency of utterances per emotion in the MELD, the MELD extended with the TEID and the resampled data.

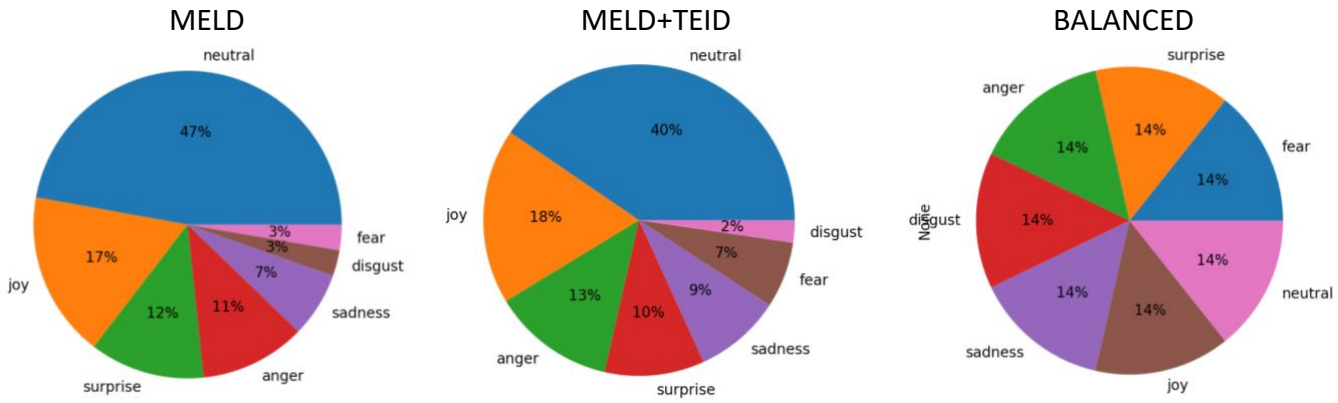

Table 9: Distribution of the emotions in the MELD, the MELD extended with the TEID and the resampled data.

Appendix 2. Overview of the models

Emotion classification models

Settings	**1**	**2**	**3**	**4**	**5**
Data	MELD	MELD	MELD	MELD	MELD+TEID
Balanced	False	True	True	True	True
Classifier	SVM	SVM	SVM	SVM	SVM
Repres.	Embed.	Embed.	Embed.	Embed.	Embed.
Embedding model	glove-wiki-gigaword	glove-wiki-gigaword	word2vec-google-news	glove-twitter	glove-wiki-gigaword
Dimensions	300	300	300	200	300
Stopwords	True	True	True	True	True
Freq. thres.	4	4	4	4	4

Table 10: The settings of the 5 emotion classification models

Topic classification models

Settings	**1**	**2**	**3**	**4**	**5**
Model	Embed.	Wordnet	Embed. + Wordnet	Embed. + Wordnet	Embed. + Wordnet
Embedding model	glove-wiki-gigaword-300	glove-wiki-gigaword-300	glove-wiki-gigaword-300	glove-wiki-gigaword-300	glove-wiki-gigaword-300
Stopwords	True	True	True	True	True
Neighborh.	10	10	10	10	10
Processing	Tokenize	Tokenize	Tokenize	Lemmatize	Lemmatize
Hyponyms	-	True	True	True	True
Hypernyms	-	True	True	True	True
Similarity metric	Embeding model	Embedding model	Embedding model	Embedding model	Wordnet Path
Sim. thres.	1.0	1.0	1.0	1.0	1.0

Table 11: The settings of the 5 topic classification models

Appendix 3. Responses to test utterances

Colors indicate: correct topic + emotion, correct emotion, correct topic

Test utterance	Setup 1	Setup 2	Setup 3	Setup 4
No, you don't get to do that. You don't get to pretend that loosing my puppy is nothing.	I don't think it is unreasonable to be angry about such animals	I don't think it is unreasonable to be angry about such animals	I don't think it is unreasonable to be angry about such animals	I don't think it is unreasonable to be angry about such animals
I asked you if the sauce was spicy and you said no, why did you lie?	I see	I see	I read that your expectations have been exceeded, why actually?	I read that your expectations have been exceeded, why actually?
See! This is just the kind of thing that gets you fired.	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.
I am not getting into the field until he leaves the grounds	Sport is just not fair sometimes.	It seems emotionally difficult to be there.	Sport is just not fair sometimes.	It seems emotionally difficult to be there.
So what's this guy's deal? Does he smell like a rodent?	I am experiencing some negative vibes, could you tell me more about it?	Such people should not live on this earth.	I am experiencing some negative vibes, could you tell me more about it?	Such people should not live on this earth.
Why don't you take that roten puddin with you, I am never going to try it	That grips my heart, maybe you can explain your story even further.	That grips my heart, maybe you can explain your story even further.	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.
I have a bad hitory with that guy.	I am experiencing some negative vibes, could you tell me more about it?	Such people should not live on this earth.	That grips my heart, maybe you can explain your story even further.	I find it very touching to hear.
Do you want me to help you? I can grab your sportsbag with your smelly shoes	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?
Ah, Grace! There's a rat!	That cheers me up too! Do you have more to say about that?	That cheers me up too! Do you have more to say about that?	That cheers me up too! Do you have more to say about that?	That cheers me up too! Do you have more to say about that?
Well I am sorry if I scared you, but the truth is your little rat scared me!	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?
I already called all the del Taco's around, I'm afraid we can't find her	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.
You told me you felt safe if I felt safe, but I don't feel safe if you are friends with those kids	Can you indicate what exactly makes you so anxious?	What a terrifying story... I am all shaking.	I am experiencing some negative vibes, could you tell me more about it?	Such people should not live on this earth.
I need you to stay where you are and call 911. Someone is in the house	That's how it goes there indeed.	That's how it goes there indeed.	That's how it goes there indeed.	That's how it goes there indeed.
We're in an elevator that is temporarly standing still.	I am experiencing some negative vibes, could you tell me more about it?	I am experiencing some negative vibes, could you tell me more about it?	I am experiencing some negative vibes, could you tell me more about it?	I am experiencing some negative vibes, could you tell me more about it?
Nobody adores panda bears as much as I do	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?	That cheers me up too! Do you have more to say about that?	That cheers me up too! Do you have more to say about that?
Let's go outand have amazing dinner together	Such people make me happy!	Such people make me happy!	Such people make me happy!	Such people make me happy!
I'm going to make you a big, fat smoothie.	I see	I see	I am experiencing some negative vibes, could you tell me more about it?	I am experiencing some negative vibes, could you tell me more about it?
You were always a good boy, now I know you are a good man	Such people make me happy!	Such people make me happy!	Such people make me happy!	Such people make me happy!
I go to a restful place in my mind and I let the'painting come to me.	That's how humans are.	That's how it goes there indeed.	That's how humans are.	That's how it goes there indeed.
It's where I first knew that he felt what I felt.	Can you indicate what exactly makes you so anxious?	Can you indicate what exactly makes you so anxious?	That grips my heart, maybe you can explain your story even further.	That grips my heart, maybe you can explain your story even further.

I mean that's silly, Of course bassets cannot see colors	I see	I see	I see	I see
The parrot is quiet and he brought his own food.	Ew, that's gross.	Ew, that's gross.	Ew, that's gross.	Ew, that's gross.
You seem to like drinking alone	Ew, that's gross.	Ew, that's gross.	Ew, that's gross.	Ew, that's gross.
She works in my doctor's office. She isn't a doctor or a nurse, but she bosses them all around.	That's how it goes there indeed.	That's how it goes there indeed.	I don't like places like that either.	I don't like places like that either.
He was so shock. So hurt. And then so furious.	I feel your anger, mate. Go ahead and tell more about it.	I feel your anger, mate. Go ahead and tell more about it.	I see	I see
It happens to be our house	I don't like places like that either.	I don't like places like that either.	I don't like places like that either.	I don't like places like that either.
Emphasize the view, the neighbourhood, the excellent school district.	That's how it goes there indeed.	That's how it goes there indeed.	I don't like places like that either.	I don't like places like that either.
I'm switching the channel to watch yesterday's game	I agree it can be scary to be in such a situation.	I agree it can be scary to be in such a situation.	Sport is just not fair sometimes.	Sport is just not fair sometimes.
When you left, I thought my life was over. I can't sleep in that big bed without him.	That's a sad story.	That grips my heart, maybe you can explain your story even further.	Ew, that's gross.	I am experiencing some negative vibes, could you tell me more about it?
All I know is that I feel horrible inside for pretending to be someone that I wasn't for all these years	Can you indicate what exactly makes you so anxious?	What a terrifying story... I am all shaking.	That grips my heart, maybe you can explain your story even further.	I find it very touching to hear.
Once I visit the beach house, I will have nothing left to live for	People like that make me angry too.	People like that make me angry too.	Such people make me happy!	Such people make me happy!
I'm not without cheer. I'm a little down about tonight's game. It's an honest down, and since you are so up, there's a homeostatic balance.	Sounds like talent!	Sounds like talent!	Sounds like talent!	Sounds like talent!
Why do we have to make such a big deal of this? I just do not like reptiles	I read that your expectations have been exceeded, why actually?	You kidding me? What an animal!	I am experiencing some negative vibes, could you tell me more about it?	I don't like such animals either.
Dad, did you just put the whole stick of butter in?	I am experiencing some negative vibes, could you tell me more about it?	I don't like places like that either.	That grips my heart, maybe you can explain your story even further.	It seems emotionally difficult to be there.
And you came here, to our house. To her house.	I don't like places like that either.	I don't like places like that either.	That's how it goes there indeed.	That's how it goes there indeed.
I played by all the rules! Why didn't you tell me there weren't any rules, it's not fair!	I agree it can be scary to be in such a situation.	I agree it can be scary to be in such a situation.	Sport is a world of emotions, anger is one of them.	Sport is a world of emotions, anger is one of them.

Table 12: The responses of the four chatbot systems to the test utterances