Chatbot report HLT

By Jona Bosman

Contents

Introduction

This report shows and discusses the results of four different systems on topic detection and emotion classification. The four systems were created by using combinations of two different topic detectors and two different emotion detectors and were evaluated on how well they could detect the topics and emotions present in a test set consisting of sentences. The systems were designed as a chatbot that generated responses to the sentences in the test set, based on the detected topic and emotion. For evaluation, the responses of the systems were compared, along with the precision, recall and F-measure for all topics and emotions.

Topic detection (or topic analysis) is the task of extracting the topic from a text. It is performed by matching the tokens from a text to keywords that are associated with topics. In this work, there were five fixed topics to be detected, so it became a classification task.

Emotion classification is the task of detecting the emotion that is expressed in a text. An emotion classifier has to be trained on a database of texts along with their emotion labels. By iteration over the data, the classifier will learn which features of a text are important for certain emotions. After training, a classifier is able to classify unseen texts.

The code for this work can be found at https://github.com/JonaBenja/hlt-report.

Description of the design

The four designs that were created consist of combinations of 2 different topic detectors and 2 different emotion detectors.

WORD EMBEDDING MODELS

Pre-trained embedding models were used for both topic detectors. The first model that was used was the GloVe Twitter word embeddings by Pennington et al 2014. The data consists of 200 dimensional vectors for 27 billion tokens, belonging to 2 billion Tweets. The GloVe Twitter word embeddings can be obtained by this link: https://nlp.stanford.edu/pubs/glove.pdf.

The other model that was used was the Google News database, which has been published by Google. This data contains 300 dimensional vectors for 3 million tokens. The Google News database can be obtained by this link: https://code.google.com/archive/p/word2vec. See table 1 for the differences between the two word embedding models.

	GloVe word embeddings	**Google News embeddings**
Type of content	Tweets	News articles
Number of tokens	27 billion	3 million
Vector dimensions	200	300

Table 1: comparison of word embedding models

The two trained word embedding models were each used to enrich the test sentences for topic detection and to train 2 different emotion classifiers, of which the architecture is described below.

TOPIC DETECTION

The sentences in the test set all belonged to one of the following topics: *animals*, *people*, *sports*, *food* and *places*. A set of keywords was provided was each of these topics in one large .json file that was loaded in the detector. An example of keywords for the topic *animals*: 'dog', 'cat', 'rat', 'reptile'. To retrieve the topic of a test sentence, the words of the sentence were tried to be matched with the keywords of the topics. The topic of the first match that was found was returned.

To facilitate keyword matching, each sentence was enriched with 10 extra words. For each word in the test sentence that was present in the embedding model, the 10 words that had the most similar embeddings were added as enrichment. This made the chance of finding a match between the topic keywords and the words in the test sentence higher, and the chance of failing the topic detection lower.

The main difference between the two topic detection systems was the word embedding model that was used to enrich the test sentences. Since only words that are present in the embedding model can be enriched, the size and content of the model would determine whether or not enrichment is possible.

Most of the test sentences are about personal affairs, which is shown in sentences like: "No, you don't get to do that. You don't get to pretend that losing my puppy is nothing.", "You told me you felt safe if I felt safe, but I don't feel safe if you are friends with those kids" and "All I know is that I feel horrible inside for pretending to be someone that I wasn't for all these years". Since the GloVe embedding model is made up of Tweets and the Google News embedding model is made up of news articles, it is expected that the topic detection system that uses the GloVe model will have a better recall, since more words are shared between the test sentences and the embedding model. The

same argument holds for the size of the models, since the GloVe model has a size of 27 billions tokens and Google News has a size of 3 millions tokens.

The vector size of the embeddings could also make a difference in performance. According to Yin & Shen (2018) “a smaller vector size could lead to a word embedding with a small dimensionality is typically not expressive enough to capture all possible word relations”. A larger vector size could overfit the data used for making the word embeddings. Following Mikolov et al (2013), it is estimated that the Google News embedding model, which has a vector dimension of 300, will perform better on the topic detection task.

The topic detection system that used the GloVe embeddings will be referred to as ‘Topic 1’ and the topic detection system that used the Google News embeddings will be referred to as ‘Topic 2’.

EMOTION CLASSIFICATION

Two classifiers were trained to classify the test sentences into emotion categories. The test sentences all belonged to one of the following emotion categories: *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness* and *surprise*. The data used for training was derived from the MELD project (Poria et al, 2019) and contained 13000 utterances from the tv-series *Friends*. The dataset is available on http://affective-meld.github.io. The training set is imbalanced: 48% of the utterances were labeled as *neutral*, 15% as *joy*, 13% as *surprise*, 11% as *anger*, 8% as *sadness*, 3% as *disgust* and 2% as *fear*.

Both classifiers were linear Support Vector Machines, with 2000 iterations over the training data. They were both trained on each of the word embedding models used for the topic detection, resulting in four different designs.

The two classifiers differed in the number of times the words from the utterances had to occur in the training data (word frequency) and whether or not common stopwords of English, as they are defined in the Nature Language Toolkit (Bird et al, 2009), were excluded from the utterances during the training process.

Words that occur only in a few utterances of the training data are considered to have less value and can make a classifier more prone to overfitting, since the words are less likely to occur in other data to be classified. (Forman, 2003). This would then influence the performance of the system.

Stopwords are words that occur in a large amount of texts. They would occur in a lot of the utterances in the data and they have a low discriminative value for any class (Forman, 2003). Thus, they are argued to be “noise” in the data and they could disturb the training process.

The imbalance of the training data can result in the classifiers classifying most test sentences as neutral. It is expected that stopwords will occur in many of the training utterances, which will make the classifier learn to associate stopwords with the *neutral* emotion class.

Generalizing, one could say that only including frequent words and excluding stopwords is filtering out noise from the data. A logical hypothesis would be that the classifier that filtered the noise would improve precision for *neutral*, and improve the recall for other emotions.

TESTING

Four designs were created by combining the two topic detectors and two emotion classifier designs, (see table 2). The word embedding model used for topic detection was the same model as the model the emotion classifier was trained on. All designs were evaluated on a test set that contained 36 sentences, along with the topic and emotion labels for all sentences. Each design generated a response to the test utterance, based on the detected topic and emotion class. For each topic and emotion, a maximum number of four responses was provided, of which the design picked a random one.

	Design 1	**Design 2**	**Design 3**	**Design 4**
Topic Detection embedding model	GloVe	GloVe	Google News	Google News
Word frequency	10	1	10	1
Stopwords	Excluded	Included	Excluded	Included

Table 2: architecture of the four designs

Results

All four designs were evaluated on the test set. Recall, precision and F-measure for each design are shown in the tables below. The responses of the designs on the test set are provided in appendix A.

TOPIC DETECTION

	Topic 1			**Topic 2**		
	Recall	Precision	F-measure	Recall	Precision	F-measure

animals	0.875	1.0	0.933333	1.0	1.0	1.0
people	0.875	0.4375	0.583333	0.875	0.5	0.636364
sports	0.2	1.0	0.333333	0.4	1.0	0.571429
food	0.857143	1.0	0.923077	0.857143	1.0	0.923077
places	0.5	1.0	0.666667	0.625	1.0	0.769231
Average	0.6614286	0.8875	0.6879486	0.7514286	0.9	0.7800202

Table 3: results for the two topic detectors

EMOTION CLASSIFICATION

Both classifiers were trained on each of two the embedding models, therefore the results of all four designs have to be presented for comparison.

RECALL EMOTION CLASSIFICATION

	Design 1	**Design 2**	**Design 3**	**Design 4**
anger	0.0	0.0	0.0	0.0
disgust	0.0	0.0	0.0	0.0
fear	0.0	0.0	0.0	0.0
joy	0.0	0.0	0.0	0.0
neutral	1.0	1.0	1.0	0.625
sadness	0.0	0.0	0.0	0.0
surprise	0.0	0.25	0.0	0.0
Average	0.1428571429	0.1785714286	0.1428571429	0.08928571429

Table 4: recall for each class of the emotion classifier of all four designs

PRECISION EMOTION CLASSIFICATION

	Design 1	**Design 2**	**Design 3**	**Design 4**
anger	0.0	0.0	0.0	0.0
disgust	0.0	0.0	0.0	0.0
fear	0.0	0.0	0.0	0.0
joy	0.0	0.0	0.0	0.0
neutral	0.242424	0.2857143	0.2222222	0.1923077
sadness	0.0	0.0	0.0	0.0
surprise	0.0	0.3333333	0.0	0.0
Average	0.034632	0.08843537143	0.03174602857	0.02747252857

Table 5: precision for each class of the emotion classifier of all four designs

F-MEASURE EMOTION CLASSIFICATION

	Design 1	**Design 2**	**Design 3**	**Design 4**
anger	0.0	0.0	0.0	0.0
disgust	0.0	0.0	0.0	0.0
fear	0.0	0.0	0.0	0.0
joy	0.0	0.0	0.0	0.0
neutral	0.3902439	0.4444444	0.3636364	0.2941176
sadness	0.0	0.0	0.0	0.0
surprise	0.0	0.2857143	0.0	0.0
Average	0.05574912857	0.1043083857	0.05194805714	0.0420168

Table 6: F-measure for each class of the emotion classifier of all four designs

Conclusions & How to improve

TOPIC DETECTION

Both detectors performed best on recall for *animals*, *people* and *food*. Topic detection system 2 showed better results for recall, precision and F-measure, averaged over all topics as well as for all topics individually. This is a surprising observation. Based on content, size and vector size it was expected that topic detection system 1 would perform better on both accuracy and precision. A possible explanation could be that the vector size of 300, used by the GloVe embeddings overfitted the data and that an embedding with a vector size of 200 performs better. This could be further investigated by creating two embedding models with a different vector size from the same data.

EMOTION CLASSIFICATION

A first observation is that all four classifiers detected *neutral* for most of the test sentences. This can be explained by the imbalance of the training data; 48% of the training utterances were labeled as *neutral*.

To accurately compare the results of the classifiers, the designs need to be compared pairwise.

Design 1 and design 2 were both trained on the GloVe word embeddings, but differed in word frequency and stopwords. Both designs classified all neutral sentences correctly, but only design 2 was able to detect another emotion: *surprise*. This results in a slightly better F-measure for emotions *neutral* and *surprise* for design 2 over design 1. This would mean setting the word frequency low and the inclusion of stopwords improves the recall and precision of the emotion classifier.

This is in contrast with design 3 and 4, which were both trained on the Google News embeddings. Both designs were only able to detect *neutral* and design 3 got a slightly better recall, precision and F-measure for this class. This would mean the opposite of our previous conclusion: setting the word frequency higher and the exclusion of stopwords improves the recall and precision of the emotion classifier. The influence of changing these parameters of the emotion classifier can thus not be determined.

Next, we compare the results between design 1 and 3, which were trained on different embedding models but set the word frequency to 10 and excluded stopwords. Both designs are only able to detect *neutral*, although design 1 has a slightly higher precision for it. This also results in a slightly higher F-measure for design 1 over design 3.

Lastly, we compare design 2 and design 4, which were trained on different embedding models but set the word frequency to 1 and included stopwords. Next to *neutral*, design 2 detected *surprise* and got a higher recall, precision and F-measure for this class. Design 4 only detected *neutral* and had a lower recall, precision and F-measure for it than design 2.

Based on these findings we could say that the embedding model the emotion classifier was trained on can (slightly) influence its performance, but changing the word frequency and in- or excluding stopwords does not. More research needs to be done to confirm these claims. For example, two other classifiers could be trained on the same embedding model but with a different vector length, or a different size. Additionally, two classifiers could be trained on different embedding models with the same vector length and size, but a different content.

RESPONSES

The responses on the test set utterances were dependent on which topic was detected and which emotion was classified. A full table with all responses of all designs can be viewed in appendix A, but a relevant selection is shown below in table 7.

It is not a surprise that most responses derived from the *neutral* emotion. Design 2 was the only design that correctly classified the test sentence 'Dad, did you just put the whole stick of butter in?' (labeled as *surprise*) as *surprise*, while the other systems classified it as *neutral*. The sentence "So what's this guy's deal? Does he smell like a rodent?" (labeled as *disgust*), was classified as *surprise* by design 2, but as *neutral* by the other systems. It occurred in a few more cases that only design 2 did not classify the sentence as *neutral* and this shows that design 2 was slightly less likely to label every sentence as *neutral*. This is in contradiction with our expectations, since design 2 did not filter for stopwords and frequent words.

Design 1 and design also classified some sentences as another emotion than *neutral*, but never as the correct emotion.

Lastly, the sentence "He was so shock. So hurt. And then so furious" (labeled as *neutral*) could not be answered by design 1 and design 2, which meant that the topic could not be detected by these two systems. That could be due to the word 'shock', that seems to be misspelled, and should have been 'shocked'. This could have prevented keyword matching and thus topic detection.

Test sentence	**Design 1**	**Design 2**	**Design 3**	**Design 4**
Dad, did you just put the whole stick of butter in?	That's true	I never expected that from you.	OK	So happy I know this now.

So what's this guy's deal? Does he smell like a rodent?	I see	I'm just as shocked as you.	OK	Meh
Why don't you take that rotten pudding with you, I am never going to try it	So happy I know this now.	Ahh so it was your tears that made my food that salty!	OK	Pineapple on pizza is what makes me angry.
Ah, Grace! There's a rat!	I love animals!	I love Freek Vonk!	OK	Meh
He was so shock. So hurt. And then so furious.	I'm afraid I cannot answer that.	I'm afraid I cannot answer that.	Thanks for telling me about them.	People are fine.

Table 7: relevant selection of appendix A

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc

Forman, G. (2003). *An extensive empirical study of feature selection metrics for text classification.* Journal of machine learning research, 1289-1305.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation.* ACL https://arxiv.org/pdf/1810.02508.pdf

Yin, Z., & Shen, Y. (2018). *On the dimensionality of word embeddings.* Advances in Neural Information Processing Systems, 887-898