Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 190 (2021) 560–563

Procedia Computer Science

www.elsevier.com/locate/procedia

2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society

Recognition of emotions in verbal messages based on neural networks

Inna S. Malova, Daria V. Tikhomirova1

National Research Nuclear University MEPhI, 31 Kashirskoe shosse, Moscow, 115409, Russian Federation

Abstract

Emotion detection and recognition by text is an under-researched area of natural language processing (NLP), which can provide valuable input in various fields. Speech and Emotion Recognition (Speech Emotion Recognition SER) has potentially wide applications, such as interaction with robots, banking, call centers, car onboard systems, computer games, etc.

Keywords: audio file segmentation, emotions, convolutional neural networks, MFC, CNN, Daily dialogues, service, speech emotion recognition, deep learning, natural language.

1. Introduction

Now there are various methodologies for solving the problem of recognizing emotions in speech. The main disadvantages of some methods are the low accuracy of emotion recognition or the long duration of work. Previously presented models mainly used vector representation of words that represent a large amount of semantic and syntactic information, but these models cannot capture the emotional relationships between words. That is why it is so important to have a competent system for accurate recognition of emotions in speech, and that the minimum amount of time is spent. Despite a lot of research in this area, the problem of automatic recognition of the emotional state of speech is not currently solved. completely solved. The recognition process is a very complex task in the field of mathematical formalization, as well as in the search for ways to clearly specify emotions based on a speech signal. Basically, the solution to this problem lies in the correct selection of the training data set, as well as the correct classification of

* Corresponding author.
E-mail address: dvsulim@mail.ru

1877-0509
10.1016/j.procs.2021.06.065

emotions. The purpose of this work is to build a service that performs automatic processing of verbal messages based on neural networks and determines the speaker's emotions.

2. Data preparation and neural network design

2.1. Preparing the audio component of the message

The first step in emotion recognition is to convert the sound wave to digital format. To do this, the sound wave sampling procedure is performed. The frequency range of human speech fits into the 4kHz band, then, according to Kotelnikov's theorem, a sampling frequency of 8KHz is sufficient to restore the signal. After performing this operation, the output will be an array of numbers, each of which represents the amplitude of the sound wave at intervals of 1/8000 seconds (Fig. 1).

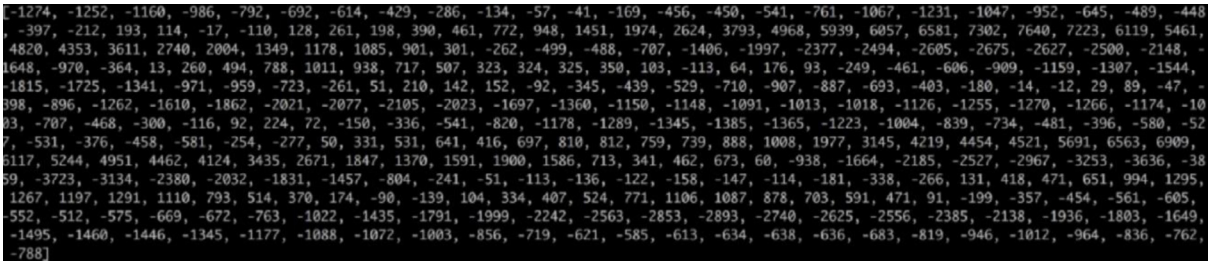

Fig. 1. *Sampled sound wave.*

You can train a neural network based on this data, but recognizing speech models by processing such a large amount of input data is difficult. You need to make the task easier by performing some preprocessing of the audio data. MFCCS (Mel - Frequency Cepstral Coefficients, Mel-frequency cepstral coefficients) are used as characteristic features of the source signal. To extract the MFCC, you need to decompose the sound wave into its individual components.

The obtained Mel-frequency cepstral coefficients can be further used as a unique characteristic of the input sound wave.

2.2. Preparing the text component of a message

At this stage, the text selected from the audio file is cleared of characters that do not affect the classification or do not carry a semantic load in this task.

Data preprocessing:

- removing all links from the text;
- removing all emoji characters from the text;
- convert all text to lowercase;
- deleting words unknown to the classifier;
- removing articles, prepositions, and conjunctions from the text;
- converting negative particles;
- splitting text into words based on punctuation marks.

2.3. Neural network architecture for text analysis

In this paper, a convolutional neural network is used for text analysis, which has the following architecture:

1. The first level of processing consists of two ultra-precise layers and max pooling with 128 filters of size 5, having an output of 216x128x2 neurons.

2. The second level of processing consists of three convolution layers also with 128 filters. At the output of the second layer, the model has 216x128x3 neurons.
3. The last level of processing is a fully connected layer that classifies 10 emotions.

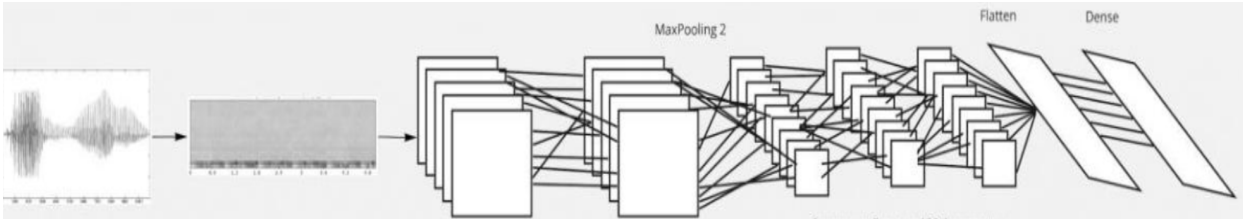

Fig. 2. Algorithm for determining emotions using a convolutional neural network.

2.4. Neural network architecture for intonation analysis

In our work, we proposed a combined neural network consisting of bilstm and CNN subnets. BiLSTM is also used for recognizing emotions in cross-language texts, which uses cross-language functions and the lexical level function to analyze texts with multilingual forms.

To include a context-sensitive word, an attention-based BiLSTM model is introduced, which helps determine the importance of each word for the task of recognizing emotions. They used three methods, such as text, emoticons, and images, to encode different information to Express emotions. The convolutional neural network (CNN) was chosen as the second subnet because CNN models demonstrate the benefits of extracting complex emotional characteristics.

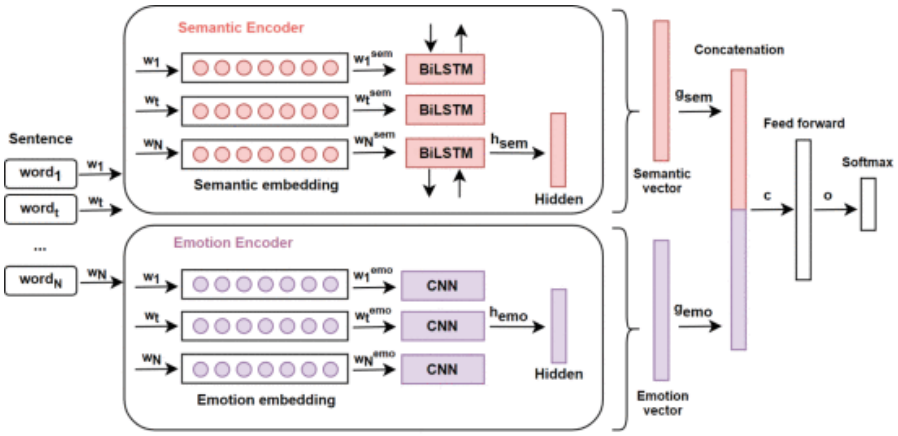

Fig. 3. Architecture combined neural network

In this paper, 3 variants of vectorization were previously considered. To select the most appropriate method, the F-measure was used. This metric allows you to quickly make a decision about the effectiveness of the neural network and the selected parameters. For clarity and selection of the optimal vectorization method, 3 methods were used:

1. Word2Vec
2. GloVe
3. FastText

Also, for comparison, we selected 4 of the most high-quality databases collected for training. The results of the experiment itself are presented in Table 1.

Table 1. F-score model parameters

Model	Vectorization method	DailyDialog	ISEAR	CrowdFlower	Emotion lines
CNN + LSTM	Word2Vec	84.2	**91.0**	55.4	60.3
	GloVe	84.3	90.4	54.6	**61.7**
	FastText	**84.8**	90.4	**56.3**	60.9

The experiment showed that the most effective combination of the CNN+LSTM model using the word2vec vectorization method and trained on the ISEAR database.

3. Conclusion

Thus, in this paper, we chose the most productive combination of the training dataset and methods for implementing the system.

It is assumed that this work will be of great practical significance, as it will simplify and automate the process of analyzing audio files and determining the speaker's emotions, which can be used in call center systems during calls.

References

[1] Vorontsov K. V. clustering and multidimensional scaling Algorithms. A course of lectures. MSU, 2017.

[2] Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 2019. Vol. 31,no. 3.

[3] Ramsundar B. TensorFlow for deep learning: Translated from English / B. Ramsundar, R. B. Zadeh // Saint Petersburg: BHV-Petersburg, 2019-256 p

[4] Zenodo [Electronic resource]: The Ryerson Audio-Visual Database of Emotional Speech and Song. - 2018. - URL: https://zenodo.org/record/1188976/?f=3#. XCSVOS1eOu4 (accessed 29.10.2020).

[5] Galushkin A. I. neyronnye SETI: osnovy teorii [Neural networks: fundamentals of theory]. —Electron. Dan. - Moscow: Hotline-Telecom, 2017. - 496 p — - access Mode:https://e.lanbook.com/book/111043.

[6] Zakovryashkin A. S., Malinin P. V., Lependin A. A. primenenie rasstroeniy Mel-chastotnykh kepstralnykh koefitsientov dlya goldos'ovoy identifikatsii lichnosti [Application of distributions of small-frequency cepstral coefficients for goldos identification of a person]. Computer engineering "

[7] Scholle F. Deep learning in python / St. Petersburg: Piter, 2019. - 400