International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 14 (2018) pp. 11512-11517

Text to Speech Conversion with Emotion Detection

Anita.R [#1], Srinivasan.R [#2]

[#]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, TamilNadu-603203, India.*

Abstract

Emotion detection from text has been a great find in the field of information recognition since it could help in the better understanding of the data and could help with complex processing methods to provide accurate results. Plain text have been used in various fields such as text mining, but the introduction of the concept of emotion detection has provided a new dimension to the field of information processing making it more efficient and providing a better scope for further enhancements of the future generations. In this paper, various complex algorithms and techniques have been utilized to implement emotion detection from text and converting the text to emotional audio. This is done by recognizing patterns of emotion and retrieving the audio files from the multimedia database.

Keywords: Text to speech conversion, Pattern recognition, Emotion detection.

INTRODUCTION

In this paper, the emotions are identified and converted into speech. This technique can be enhanced further which could be used in various fields such as robotics and automated systems which could provide a more human like interaction with the users which could improve the overall working environment and the comfortability of the users. Machines with more emotional human like voices will make it easier for the users to interact the machines and could make a lot of people attracted towards the usage of such kind of devices. For example, robots with more emotional voice outputs add an extra dimension to the field of robotics which could totally eradicate the feeling of interacting with the computational device.

In this paper, we have utilized various complex algorithms and techniques to implement emotion detection from text. This is done by recognizing patterns of emotion and utilizing them to understand the overall emotion present in the given statement. Complex statements with multiple emotions could also be used and still it is possible to identify the emotion which has more strength than the other by computing the emotion strength, by analyzing various words, the position of the word present and the overall meaning or the requirement of the word to be used in the sentence. This provides us with a more accurate emotion which has been present in the given sentence and it cannot break down further to be analyzed.

BACKGROUND

This section describes about the types of emotions, text to speech conversion, machine learning, supervised learning techniques, data classification techniques such as support vector machine classifier and random forests.

TYPES OF EMOTIONS

Emotion detection procedure deals with finding the emotion present in the given sentence provided as input by the user. The emotions are classified into various categories such as Happy, Sad, Angry and Fear. Any other emotional type falls under any one of these basic four emotional categories. For example frustration is an emotion that falls under the category angry. Hence any other complex emotion which is present in the given sentence falls under any one of the given four basic emotional categories. An emotion could be a combination of any of two or more basic emotions. But the algorithm analyses the given sentence and determines the emotion of the sentence based on the strength of the emotion in the given sentence. This is done by analyzing the sentence pattern and identifying the emotion which represents the given pattern of sentence.

TEXT TO SPEECH CONVERSION

Once the emotion present in the text sentence is identified by the algorithms, the given emotional sentence textual data is converted into audio file with the emotion associated with the given sentence. This is done with the help of a text to speech converter. A text to speech converter converts natural language text into audio speech. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output. Text-to-speech convention transforms linguistic information stored as data or text into speech. It is widely used in audio reading devices for blind people present days. Detecting emotional state of a person by analyzing a text document written by him/her appears challenging but also essential many times due to the fact that most of the times textual expressions are not only direct using emotion words but also result from the interpretation of the meaning of concepts and interaction of concepts which are described in the text document.

MACHINE LEARNING

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension as is the case in data mining applications machine learning uses that data to detect patterns in data and adjust program actions accordingly. Machine learning algorithms are often categorized as being supervised and unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from data-sets.

2.4 Supervised Learning Techniques

Majority of machine learning algorithms currently available are supervised machine learning approaches. In Supervised Machine learning, training data-sets and test data-sets are used. Throughout the process, it makes use of a methodology to comprehend the function mapping. The objective is to roughly decide the function mapping with the help of training model, so that when new keyword is given as input, output can be predicted for that keyword. It is known as supervised learning approach because the procedure of an algorithm learning the model from the predefined training data-set is analogous to a person overseeing the process of learning. Right answers are known and the machine learning logic consequently makes prognosticate on the training data and is rectified by the human overseeing the learning process. When it attains an adequate level of authenticity, Learning halts. In supervised learning, the output data-sets are provided which are used to train the machine and get the desired outputs whereas in unsupervised learning no data-sets are provided, instead the data is clustered into different classes.

DATA CLASSIFICATION TECHNIQUES

An algorithm that performs categorization, especially in a tangible implementation, is known as a classifier. Mathematical functions can also perform classification by implementing a classification algorithm that maps input data to a class. There are many varieties of classifiers available.

Support Vector Machine Classifier

A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification (Li and Jain 1998) and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Random Forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model.

RELATED WORK

The basic hypothesis presented in this paper is to categorize emotions into primary emotions and secondary emotions. More recently, researchers have investigated several aspects of human emotions in order to arrive at a set of emotion categories that are universally acceptable. Several works in this direction have been reported in this literature survey. It lists the basic emotion categories identified by the different researchers.

Some psychologists (Rachael et al 2014) have investigated facial expressions of emotion to identify the basic distinguishable expressions among them, and mapped them to basic human emotions. (Paul 1993) has defined basic emotions as those that have universally accepted distinctive facial expressions. A lot of research has been done in the area of emotion classification from text. The majority of work has been done on lexicon based supervised learning approaches.

Satapathy and Bhagwani (2012) worked on identifying the emotion conveyed by the sentence at different granularities like one is coarse grained classification. It classifies the sentence into only positive and negative emotions based on their polarity, other is fine grained classification. It classifies the emotion conveyed by sentence based on predefined emotion set. They also proposed a bag of word approach with lower order dependencies captured through the use of bigrams and trigrams. They work on International Survey on Emotion Antecedents and Reaction (ISEAR) dataset and WordNet. They mainly focused on normalization of text as the inflected forms of a word occur in plenty. For classification they have used Multinomial Naïve Bayes(MNB) and compared with Vector Space Classification (VSM) with MNB giving most promising result.

Danisman and Alpkocak (2008) have done emotion classification from text using Bag of Word (BOW) approach.

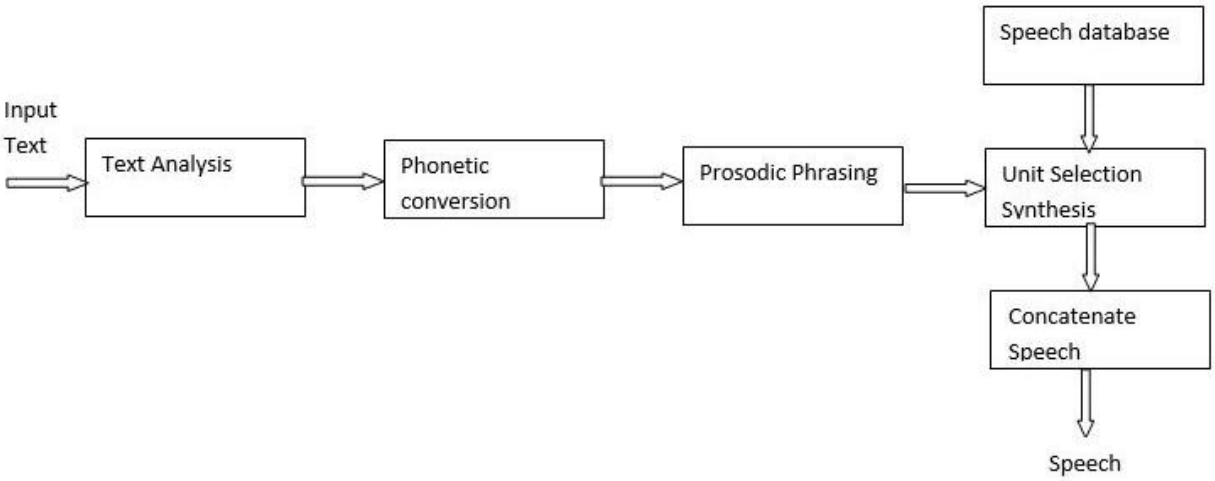

Figure 1. Architecture Diagram

ISEAR and SemEval datasets were used. They have taken 801 news headlines given by "Affective Task" in SemEval 2007 workshop. Vector Space Model has been used for the classification of emotions and valence in text. The result has been compared with other classifiers such as Naïve Bayes, Concept Net, SVM and VSM. Finally they have found that VSM classifier gave better performance than other classifiers.

Mihalcea et. al. (2006) analyzed the text for emotions automatically using unsupervised machine learning approach. They have used news headlines taken from most important newspapers such as New York Times, CNN and Google News search engine as well as from BBC news for developing dataset for emotion analysis. Two data sets have been used, one had 1000 annotated headlines and the other had 250 annotated headlines. They have used knowledge based and corpus based methods for implementing five systems for emotions analysis. The average result retrieved from the five different systems has been compared with the three SemEval System.

Li and Zong (2008) have proposed sentiment classification using multiple SVM classifiers. They have assigned the documents into some predefined category and then performed sentiment classification by the polarity of the subject. This task has been performed with the classifier combination approaches. Unigrams and some Parts of Speech (POS) features have been used to train the review data for generating different classifiers. The classifier selection can be done with the classifier selection method and they are combined with the help of some combining rules. The result has been compared with the individual classifier and concluded that combinational classifiers gave better performance than individual classifier.

Alm et. al. (2005) used supervised machine learning with SNoW (Sparse Network of Windows) learning architecture for emotion prediction on text. SNoW is used for large scale learning tasks and which is a multiclass classifier. They have recognized emotional passages and determined their valence. They have identified emotions based on emotional categories in the speech. They have used 22 fairy tales as data sets. The results have been compared with the Naïve Bayes and BOW approach, and achieved the better performance than others.

Liu and Yu (2005) have classified sentences into basic emotion categories using real-world knowledge. They have used a real world corpus of 400,000 facts. They have combined four linguistic models. This system is verified in an email writing application. The results proved that the system's robustness.

Kelemen et. al. (2004) and Lewis (1998) have used Bayesian Classifier for text categorization. The class specific features have been used as the feature sets. Bayesian classification approach for classification of texts automatically uses the class-specific attributes. The approach chooses a particular attribute subgroup for each group. Bayesian Classification rule based on PDF Projection Theorem is then constructed. It takes dependencies among attributes into account. Their results demonstrate the effectiveness of their approach in real world Data Mining usage scenarios.

PROPOSED WORK

The text input is analysed to determine the emotions. The phonetic conversion is performed and prosodic phrasing is constructed. The audio files of individual words are taken from the speech database and they are concatenated to generate the continuous speech.

The user inputs simple textual data into the system. The system identifies the emotion present in the textual data by performing several activities such as keyword identification, symbols (!,?) etc. Once the emotion is identified by the system, the user is notified about the emotion present in the textual data. This comprises of the first part of the problem. Once the emotion is detected, grammar of the sentence is identified. The text to speech conversion system analyses the textual data word by word, and selects the audio of the word.

The proposed work makes use of database which has all the words which are allowed to use on the sentence. This database is a multimedia database which utilizes and comprises of

audio files of emotions associated with the particular word. A word could be associated with more than one emotion. Hence emotional audio for a word for all possible emotions are uploaded along with the primary emotion of the word. This is stored in a multimedia database which could be recovered whenever it is required. Once the sentence with an emotional word is given as input to the system, the system first identifies the emotion which lies in the given sentence. This is done by analyzing the given words in the input sentence and identifying the sentence pattern and the strength of the words in the given sentence pattern. Once it is completed the overall emotion present in the sentence could be identified. Then the text to speech conversion process could be started. The given words are compared with the words present in the multimedia database and the matching emotional audio for the given words are retrieved from the multimedia database. The audio is retrieved based on the overall emotion of the sentence. Then all the audio files are concatenated to form a single flawless continuous audio file to produce the speech output for the user.

METHODOLOGY

This section describes data collection, pattern recognition, output audio generation and validation.

Data Collection

A single data-set consists of attributes of the word such as word length, word emotion, word audio etc. At first, all the required data for the algorithm to utilize and the data range that are to be processed by the system are collected and uploaded into the multimedia database so that they could be retrieved and processed later. If the required data is not available in the database, it will lead to error in the output since the system does have the required data to be processed and produce the output.

Pattern Recognition

Once the emotion has been determined from the given sentence, the type of sentence given as input is identified. Different type of sentences have different patterns and hence it is not possible to identify emotions from a given sentence pattern. Hence various possible types of sentences and their patterns, possibility of words occurring in the sentence, position of the word, the emotion referring the word, the strength of the word in the given place and pattern of the sentence are defined into the algorithm. So that the given emotional word occurring in any type of sentence and any place in the sentence still could be identified and the emotion could be recognized. This makes the algorithm very efficient with dealing various kinds of input data since it could deal with a wide range of emotions and words making it diverse in nature. This is a very important part of emotion detection from a given sentence.

Output Audio Generation

Once the emotion is detected, the textual data is sent through a lexical analyzer which scans the given data with natural language processing (NLP) techniques for grammatical order of the sentence. Grammatically analyzing textual data helps in understanding whether the textual data contains contradictions, arrangement of words and helps in better understanding of the emotion in the given text. This helps us to provide an emotion constant to each word in the textual data. Each word is assigned a number which indicates the emotion of that word. Now the textual data along with the allocated numbers is sent to the TTS conversion system.

The TTS conversion system analyses the textual data word by word, and selects the audio of the word by matching with the index of the word in the audio database. Each word in the database has a set of different emotional audio which can be identified with the help of numbers assigned to that audio file. Hence the correct emotional audio for the textual data is selected by comparing the two numbers. Then the audio files are arranged in the order of the textual data entered. Finally the audio files are merged in such a way that they are heard as a sentence in flow.

Validation

Validation phase is the testing of constructed model by inserting the data-sets into it for classification. If the training is done in the right way, it should work for the new cases as well as the one that is trained on by the admin. Otherwise, the training has to be done again until the required accuracy is being achieved and the attributes are mapped correctly to the class.

EVALUATION

This section discusses the performance analysis of the proposed system. The performance of the given system that is used for categorizing words based on emotion to convert them into speech audio based on its overall sentence emotion is analyzed. Here, training data is taken that consists of attributes. The attributes are the word length, word emotion and the word audio. The model is built from the training data-sets and applied on the test data to gauge the accuracy. The accuracy calculations were performed for the amount of training data ranging from 200 data-sets all the way up-to 20,000 data-sets.

Table 1. Accuracy Level in Tabular Format

Number of Training Data	Number of Test Data	Correct	Incorrect	Accuracy Level
200	20	20	0	100
500	20	20	0	100
3000	20	19	1	95
10000	20	20	0	100
20000	20	18	2	90

From the above graph, it is inferred that given algorithm, on average is correctly recognizes the data 95% of the time. It is also observed that when large number of training data is available, more accurate the predictions are, even beyond 10,000 data-sets, achieving 100% accuracy on recognition is possible and is observed.

Admin data are the data which are used by the system to process the given sentence to provide the speech output. After uploading, the data-sets are displayed in tabular format. The software, builds a model by taking into account the frequency of attributes that occur in the entire collection of data-sets and calculates the class prior probability and attribute prior probability scores. From that, likelihood of attributes given the class is also calculated.

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when data-sets are separated into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, the effects of data discrepancies can be minimized and better understand the characteristics of the model. After a model has been processed by using the training set, the model is tested by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that is to be predicted, it is easy to determine whether the models guesses are correct.

The user enters the data for which the predictions have to be done. In this application, the user has to enter the emotional word in a sentence and upload it as input. Here, the user of the application must upload their data-sets containing the word attributes. Emotions with the highest strength scores are set as target class for the data-set. It is imperative to separate the documents into training and test data. When the objective turns to prediction, and in particular towards the development of predictive models, models are used to guide many decisions, and to make hundreds, thousands, or even billions of predictions. With a predictive model the principal focus is no longer on the data but on a type of theory about reality. If the predictive models are developed, then there must be a way to assess their accuracy, reliability and credibility.

If the assumptions are more or less correct then the data that is available today is a reasonable representation of the data that is expected to have in the future. Holding back some of today's data for testing is therefore a fair approximation to having future data for testing. The division of the data into learn and test must be executed carefully to avoid introducing any systematic differences between learn and test.

The user can view the strength scores computed by the algorithm for each word. After computing, it displays score of words for each data-set that is available in the test data file. In statistics, the posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence or background is taken into account. The emotion with the highest strength is the outcome of sentence prediction.

The intricate details of the Data-set such has the length of the word, emotion associated with the word, strength value of emotion associated with word of each attribute are displayed in the form of tabular format. The frequency count of terms help to get a rough idea as to how many terms are there in the sample that gives enough idea of how many times a particular term occurs. The frequency of target class talks about accuracy and recall and mentions how close it is to the actual desired results. With the help of these things it is made sure that data is accurate and reliable. The probability scores for each attribute tell the chances of obtaining classes close to the desired results. The value should be between 0 and 1 and once the probability is found out it can be calculated accordingly.

User Interface is the medium through which the user and a computer system can interact, in particular the use of input devices and software. In this system, we have utilized the visual c# as the user medium which provides the user with easy accessibility options to utilize the system efficiently.

CONCLUSION

In this proposed system, it is possible to handle any type of complex sentences which could implement complex patterns and multiple emotional words and patterns and the system is still capable of handling the resources efficiently and produces the results accurately and in a less time consuming manner. Hence it could be used as a trust worthy text to speech conversion system along with apt emotional detection features which could handle complex emotions with ease.

In the information age, people are living a life where information needs to be available to users in the best possible manner which attracts the user but at the same time provides the users with accuracy and comfortability since there is no room for compromise. As a future work this research can be extended by adding search feature to other systems as well. This would lever-age the power of text mining to robotics to the next level of computing outputs. In addition to this, various different types of voices could be added so as to match with the requirements of the user and also to provide an alternate option for the user. These could also be used for blinds to build various voice related systems such as emotional story tellers, etc which could provide a new boost into the field.

REFERENCES

[1] Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4), 384–392.

[2] Rachael E. Jack, Oliver G.B. Garrod, Philippe G. Schyns , "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time." Current Biology, Volume 24, Issue 2, p187–192, 20 January 2014.

[3] Kelemen, A., Zhou, H., Lawhead, P., and Liang, Y. (2004). "Naive bayesian classifier for microarray data." Proceedings of the International Joint Conference on Neural Networks.

[4] Lewis, D. D. (1998). "Naive (bayes) at forty: The independence assumption in information retrieval." Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 1398.

[5] Li, Y. H. and Jain, A. K. (1998). "Classification of text documents." Comput J, 41, 537–546.

[6] Liu, H. and Yu, L. (2005). "Toward integrating feature selection algorithms for classi-fication and clustering." IEEE Transactions on Knowledge and Data Engineering, 17, 491–502.

[7] Satapathy, S., Bhagwani, S., (2012) "Capturing Emotions in Sentences". Vol 15, from http://202.3.77.10/users/sranjans/emotionDetection.pdf.

[8] Danisman, T., Alpkocak, A., (2008) "Emotion classification of text using vector space model". Springer Berlin Heidelberg, International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, 205-216.

[9] Mihalcea, R., Corley, C., Strapparava, C., (2006) "Corpus-based and knowledge-based measures of text semantic similarity". AAAI, vol 6, pp 775-780.

[10] Li, S., Zong, C., (2008) "Multi-domain sentiment classification". Association for Computational Linguistics, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, pp 257-260.

[11] Alm, C., -O., Roth, D., Sproat, R., (2005) "Emotions from text: machine learning for text-based emotion prediction". Association for Computational Linguistics, Proceedings of the conference on human language technology and empirical methods in natural language processing, pp 579-586.