International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 11 • Issue 2 pp. 64-69 April 2020 - Sept 2020 www.csjournals.com

Statistical Modeling in Sentiment Analysis using Wod2Vec

Shubhnandan S. Jamwal
Department of Computer Science and IT, University of Jammu

Abstract: Sentiment analysis in audio using automatic speech recognition is an emerging research area where sentiments exhibited by a speaker are detected from speech and many times they are also converted to text. It is considered as a relatively underexplored area of research as compared to text based sentiment detection and detection. The process of extracting sentiments of the speaker from natural audio is a challenging problem. Generic methods for sentiment extraction generally use transcripts from a speech recognition system, and process the transcript using text-based sentiment classifiers. In this paper, the concept of sentiment analyses is done on the small data set of Dogri, in order to analyze Wave2vec model because the Dogri language is highly tonal and inflectional. We have achieved a fair amount of accuracy with the data in the uncontrolled environment.

Keywords
Sentiment, Dogri, Wave2Vec, Word embeddings, Automatic Speech Recognition, Speech Data, Statistical Modeling

Introduction

Extracting sentiment automatically from speech data is a challenging problem in research. Now the consumers are making decisions and reviewing the products with the audio and producers improving their products. Wave2Vec is a model is a machine learning model using unsupervised and supervised techniques for training the machine. It revolutionized how machines can process and understand text and audio data without the need for extensive labeled datasets. Wave2Vec can be fine-tuned for ASR tasks, producing models that understand spoken language and achieves this by learning directly from raw audio waveforms and then fine-tuning for tasks like automatic speech recognition (ASR). In some cases the representations learned by Wave2Vec can be used for tasks beyond ASR, such as speaker identification or emotion recognition.

A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. Dogri language is categorized into one of the highly tonal language and, therefore, the word spoken in the data is usually varies from speaker to speaker. This is intrinsically difficult because of the curse of dimensionality: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training.

Literature Review

It is observed form the literature that the approaches based on n-grams obtain generalization by concatenating very short overlapping sequences seen in the training set. Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin [1] proposed distributed representation for words which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences. The model learns distributed representation for each word along with the probability function for word sequences, expressed in terms of these

representations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar to words forming an already seen sentence. Training such large models within a reasonable time is itself a significant challenge. They conducted experiments using neural networks for the probability function, showing on two text corpora that the proposed approach significantly improves on state-of-the-art n-gram models, and that the proposed approach allows to take advantage of longer contexts. N. Hadiya and N. Nanavati [2] discussed various available lexicon resources that are used for sentiment analysis in some Indian languages and presented the theoretical parametric evaluation of the studied techniques. They also further discussed challenges, which were identified during SA in Indian Languages. B. Gupta et al. [3] studied transfer learning and implemented it for Sentiment Analysis of Tweets by using the knowledge of Yelp reviews and observed that transfer learning approach is faster than the conventional machine learning approach and give comparable accuracy at much smaller dataset. Y. Sharma and et al., [4] extracted sentiments by using vector representation of the words using unsupervised technique. Singh, Avinash and others [5] have also carried out research on very low resourced languages like Dogri and used statistical machine translation system to train translation models for Dogri English language pair.

Speech recognition and artificial intelligence powers the automatic speech recognition systems, these systems can be applied in the call center environments. R. R. Sehgal, S. Agarwal and G. Raj [6] explain the use of sentiment analysis to identify if the customer is satisfied the ASR system's performance. They presented approaches and techniques for how sentiment analysis can be used in call centre environments to recognize user emotions. L. Kaushik, A. Sangwan and J.

H. L. Hansen [7] showed that baseline system is suboptimal for audio sentiment extraction. Alternatively, new architecture using keyword spotting (KWS) is proposed for sentiment detection. In the new architecture, a text-based sentiment classifier is utilized to automatically determine the most useful and discriminative sentiment-bearing keyword terms, which are then used as a term list for KWS. In order to obtain a compact yet discriminative sentiment term list, iterative feature optimization for maximum entropy sentiment model is proposed to reduce model complexity while maintaining effective classification accuracy. A new hybrid ME-KWS joint scoring methodology is developed to model both text and audio based parameters in a single integrated formulation. For evaluation, they developed two new databases for audio based sentiment detection, namely, YouTube sentiment database and another newly developed corpus called UT-Opinion Opinion audio archive. These databases contain naturalistic opinionated audio collected in real-world conditions. The proposed solution is evaluated on audio obtained from videos in youtube.com and UT-Opinion corpus. Their experimental results show that the proposed KWS based system significantly outperforms the traditional ASR architecture in detecting sentiment for challenging practical tasks. L. Kaushik, A. Sangwan and J. H. L. Hansen [8] proposed a system for automatic sentiment detection in natural audio streams such as those found in YouTube. The proposed technique uses POS (part of speech) tagging and Maximum Entropy modeling (ME) to develop a text-based sentiment detection model. Additionally, they proposed a tuning technique which dramatically reduces the number of model parameters in ME while retaining classification capability. Finally, using decoded ASR (automatic speech recognition) transcripts and the ME sentiment model, the proposed system is

able to estimate the sentiment in the YouTube video. In our experimental evaluation, we obtain encouraging classification accuracy given the challenging nature of the data. L. Kaushik, A. Sangwan and J. H. L. Hansen [9] also extracted speaker sentiment from natural audio streams such as YouTube. A number of factors contribute to the task difficulty, namely, Automatic Speech Recognition (ASR) of spontaneous speech, unknown background environments, variable source and channel characteristics, accents, diverse topics, etc. They proposed several enhancements including (i) better text-based sentiment model due to training on larger and more diverse dataset, (ii) an iterative scheme to reduce sentiment model complexity with minimal impact on performance accuracy, (iii) better speech recognition due to superior acoustic modeling and focused (domain dependent) vocabulary/language models, and (iv) a larger evaluation dataset. Collectively, our enhancements provide an absolute 10% improvement over our previous system in terms of sentiment detection accuracy.

M. S. Barakat, C. H. Ritz and D. A. Stirling [10] conducted experiment on user generated video product reviews in social media which is gaining popularity every day due to its creditability and the broad evaluation context. They investigated the feasibility of sentiment detection temporally from those videos by analyzing the transcription generated by a speech recognition system which was not investigated before. Another two main contribution for this paper is introducing a solution to the problem of fixed threshold estimation for the Naïve Bayesian classifier output probabilities and irrelative text filtering for improving the sentiment classification. Various experiments indicated the proposed system can achieve an F-score of 0.66 which is promising knowing that the sentiment classifier offers an F-score of 0.78 provided that the input text is error free. Thien Khai Tran [11] used the machine learning approaches to classify hotel service reviews which were integrated into voice server system. Because of this system the users can consult hotel information via phone calls instead of the keyboard. T. Chatchaithanawat and P. Pugsee [12] showed the proposed framework which enables users to know what the review about the laptop is. The objective of the research is to analyze messages from community web sites and identify subjective paragraphs and the sentiment of texts. There are three main procedures of this framework. Firstly, subjective paragraphs will be detected from the source materials by detecting subjective words. Secondly, the aspects of subjective paragraphs will be defined by the frequency of words in each aspect. Finally, the sentiment of texts is classified by the machine learning. In conclusion, the results of this framework are identified subjective paragraphs in each aspect and the sentiment of texts in paragraphs. So this technique framework is useful for the system applied for analyzing the laptop reviews as to help the customer make decision before purchasing. J. Sun et al. [13] proposed a method to predict the user emotional state (anger or neutral) for improvement of user satisfaction in call center. They have done a series of experiments on human-human dialogues which derived from China Mobile call center corpus, and system gives each dialogue a user emotion result that is angry or neutral, corresponding to the user satisfaction or dissatisfaction. Deepak Baby, Tuomas Virtanen, Jort F. Gemmeke, and Hugo Van hamme [14] proposed an efficient way to directly compute the full-resolution frequency estimates of speech and noise using coupled dictionaries: an input dictionary containing atoms from the desired exemplar space to obtain the decomposition and a coupled output dictionary containing exemplars from the full-resolution frequency domain. They introduced modulation spectrogram features for the exemplar-based tasks using this approach. The proposed system was evaluated for various choices of input exemplars and yielded improved speech enhancement performances on the AURORA-2 and AURORA-4 databases. Kuan-Yu

Chen, Shih-Hung Liu, Berlin Chen, Hsin-Min Wang, and Hsin-Hsi Chen [15] presented a novel method for learning the word representations, which not only inherits the advantages of classic word embedding methods but also offers a clearer and more rigorous interpretation of the learned word representations. Built upon the proposed word embedding method, they formulated a translation-based language modeling framework for the extractive speech summarization task. A series of empirical evaluations demonstrate the effectiveness of the proposed word representation learning and language modeling techniques in extractive speech summarization.

Modeling Wave2Vec for Dogri Language

Wave2Vec learns general representations from raw speech data in an unsupervised manner, which means it doesn't require transcriptions or labeled data. The model takes raw waveforms as input and learns to compress these into latent representations that capture important characteristics of the speech signal, like phonetic and linguistic features. After extracting features, the representations are quantized into discrete units. This helps the model to form meaningful chunks of data that can be used for downstream tasks. Wave2Vec 2.0 is used to convert raw audio signals into meaningful speech representations in the form of embeddings. These embeddings capture the linguistic information from the speech signal, which includes phonetic and prosodic features (tone, pitch, etc.). Sentiment in speech is often conveyed not only by word choice but also by how it's said (intonation, stress, rhythm, pitch). The prosodic features, like tone, volume, and pacing, are critical in identifying the sentiment behind the spoken words.

The experiment is conducted for training the Wave2Vec in analyzing the sentiment in the text of the Dogri language. The text is classified into three categories that is Positive, Negative, and Unknown State. The training data is composed of 235 samples and test data is composed of 40 samples. The most widely used metric for ASR performance is the Word Error Rate (WER). It measures the percentage of words that were incorrectly transcribed by the ASR system compared to a reference transcript.WER is calculated using the formula:

$$WER = \frac{S + D + I}{N}$$

S = Number of substitutions (words that are incorrectly recognized)
D = Number of deletions (words that are missed)
I = Number of insertions (extra words that are incorrectly added)
N = Total number of words in the reference transcript

Results and Discussions

The features learned from Wave2Vec are passed through sentiment classification layers (e.g., dense layers with softmax for categorical sentiment). The final testing was done by taking raw speech as input, converting it into a feature-rich latent space using Wave2Vec, and then using those features to classify the sentiment behind the speech using a trained sentiment classification model. We have achieved the WER of 21% with the amount of the data which has been used for training and testing the speech analysis of the data. ASR performance degrades significantly in the presence of background noise, overlapping speech, or low-quality recordings.

We have used the uncontrolled environment for the recordings of the data and Noise-cancellation techniques and better acoustic models can help reduce WER in these conditions.

Conclusion

Wave2Vec can capture prosodic features (intonation, stress, etc.) that are crucial in understanding the sentiment in spoken language. Wave2Vec can be pre-trained on large amounts of unlabeled speech data, making it easier to build speech sentiment systems with less labeled sentiment data. Fine-tuning Wave2Vec for sentiment analysis helps leverage the general speech representations learned during pre-training, improving accuracy on smaller datasets. This model has been instrumental in improving the efficiency of speech recognition systems, reducing the need for large annotated datasets. The uncontrolled environment for the recordings of the speech data must be used for better results. Before performing the experiments the noise-cancellation techniques must also be used for better results.

References

[1] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model", Journal of machine learning research, vol. 3, pp. 1137-1155, Feb 2003.
[2] N. Hadiya and N. Nanavati, "Indic SentiReview: Natural Language Processing based Sentiment Analysis on major Indian Languages," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 322-327, doi: 10.1109/ICCMC.2019.8819786.
[3] B. Gupta et al., "Cross domain sentiment analysis using transfer learning," 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), Peradeniya, 2017, pp. 1-5, doi: 10.1109/ICIINFS.2017.8300363.
[4] Y. Sharma, G. Agrawal, P. Jain and T. Kumar, "Vector representation of words for sentiment analysis using GloVe," 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2017, pp. 279-284, doi: 10.1109/INTELCCT.2017.8324059.
[5] Singh, Avinash & Kour, Asmeet & Jamwal, Shubhnandan. (2016). English-Dogri Translation System using MOSES. Circulation in Computer Science. 1. 45-49. 10.22632/ccs-2016-251-25.
[6] R. R. Sehgal, S. Agarwal and G. Raj, "Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 213-218, doi: 10.1109/ICACCE.2018.8441741.
[7] L. Kaushik, A. Sangwan and J. H. L. Hansen, "Automatic Sentiment Detection in Naturalistic Audio," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 8, pp. 1668-1679, Aug. 2017, doi: 10.1109/TASLP.2017.2678164.
[8] L. Kaushik, A. Sangwan and J. H. L. Hansen, "Sentiment extraction from natural audio streams," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 8485-8489, doi: 10.1109/ICASSP.2013.6639321.
[9] L. Kaushik, A. Sangwan and J. H. L. Hansen, "Automatic sentiment extraction from YouTube videos," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 2013, pp. 239-244, doi: 10.1109/ASRU.2013.6707736.

[10] M. S. Barakat, C. H. Ritz and D. A. Stirling, "Temporal sentiment detection for user generated video product reviews," 2013 13th International Symposium on Communications and Information Technologies (ISCIT), Surat Thani, Thailand, 2013, pp. 580-584, doi: 10.1109/ISCIT.2013.6645925.
[11] Thien Khai Tran, "SentiVoice - a system for querying hotel service reviews via phone," The 2015 IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF), Can Tho, Vietnam, 2015, pp. 65-70, doi: 10.1109/RIVF.2015.7049876.
[12] T. Chatchaithanawat and P. Pugsee, "A framework for laptop review analysis," 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Chonburi, Thailand, 2015, pp. 1-5, doi: 10.1109/ICAICTA.2015.7335358.
[13] J. Sun et al., "Information Fusion in Automatic User Satisfaction Analysis in Call Center," 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2016, pp. 425-428, doi: 10.1109/IHMSC.2016.49.
[14] Deepak Baby, Tuomas Virtanen, Jort F. Gemmeke, and Hugo Van hamme. 2015. Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 23, 11 (November 2015), 1788–1799. https://doi.org/10.1109/TASLP.2015.2450491
[15] Kuan-Yu Chen, Shih-Hung Liu, Berlin Chen, Hsin-Min Wang, and Hsin-Hsi Chen. 2016. Novel Word Embedding and Translation-based Language Modeling for Extractive Speech Summarization. In Proceedings of the 24th ACM international conference on Multimedia (MM '16). Association for Computing Machinery, New York, NY, USA, 377–381. https://doi.org/10.1145/2964284.2967246