WASP Project Course 2021

The efficient podcast listener:

Enhancing the Spotify podcast dataset using both audio and text

Background

Podcasts are a medium for thoughts and ideas in which different topics are debated and discussed. In these, the speakers often share emotions and use expressions such as anger, joy and excitement, but also use ways of expressing themselves, such as sarcasm or irony. These are often defined not only by what is being said, but also how. The audio is therefore a necessary tool to analyze the emotions and mood of the podcast, and this information can therefore be used in the analysis of podcasts. Furthermore, this can in a later stage enhance the recommendation of podcasts for users.

This project will make use of the Spotify Podcasts Dataset[1], consisting of 100,000 episodes from different podcasts, so which both the raw audio and automated transcriptions are available. These two modalities have the potential to provide complimentary information to detect the aforementioned categories. However, labels for sentiments for the podcasts are currently not available. Thus, the project will involve investigating lightly supervised multimodal methods to jointly leverage the text and audio data to obtain these labels.

As a starting point, we will explore the use of Edyson[2], a tool for efficiently labelling broad strokes of large audio corpora. Subsequently, we will strengthen the confidence of the annotations through the use of other semi- or lightly supervised methods.

Participants

Industrial partner: Spotify

Industrial supervisor: Jussi Karlgren, Spotify, jussi@lingvi.st, jkarlgren@spotify.com

Academic supervisor: Gustav Eje Henter, ghe@kth.se, KTH Royal Institute of Technology

Coordinating WARA representative: Jussi Karlgren, WARA-Media

Suggested WASP PhD students: Tobias Norlund, Lovisa Hagström, Ulme Wennberg, Filip Cornell and Shivam Mehta.

Challenges to investigate

- How do we create, starting from only transcripts and audio, obtain accurate annotations for podcast segments?
- How do we account for the transcriptions and the labels being imperfect?

- How do we enhance today’s text-based sentiment analysis models by using audio as an additional modality?

Resources

- The Spotify podcast dataset (information about the dataset can be found here). The dataset contains about 100,000 podcasts filtered to contain only documents which the creator tags as being in the English language.
    - Each episode in the dataset contains an audio file, a text transcript, and some associated metadata.

Deliverables

- A set of new annotations of different sentiments for the Spotify Podcast dataset over time.
- A sentiment analysis model capable of analyzing the sentiment in a podcast using both audio and text.
- A model able to segment a podcast into different moods and sentiment (not only a single label for the whole episodes).

References

1. Jones, Rosie, et al. "TREC 2020 Podcasts Track Overview." arXiv preprint arXiv:2103.15953 (2021).
2. Fallgren, Per, Zofia Malisz, and Jens Edlund. "How to Annotate 100 Hours in 45 Minutes." Proc. Interspeech. 2019. Code available at https://github.com/perfall/Edyson

Keywords

Machine learning for audio, natural language processing, sentiment analysis, audio signal processing, acoustic analysis, paralinguistics, data mining