Machine learning - ML4Science
Explaining venture teams' opportunity identification through multimodal data

Oussama Gabouj, Ahmed Aziz Ben Haj Hmida, Erwann de Belloy
supervised by Davide Bavato
EPF Lausanne, Switzerland

***Abstract*—Evaluating a team's potential is vital for discovering promising new startups. In this pursuit, this work consists into analysing the synergy within the founding team. Our objective is to harness machine learning methods [1] to derive MBTI personalities and emotions of speakers during entrepreneurial discussions. This approach aims to enhance our ability to identify the most promising teams for creating successful startups.**

I. Introduction

In this project, our focus revolves around employing machine learning algorithms to forecast the number of proficient ideas a team can generate. We achieve this by extracting features from multimodal data comprising transcripts and audio recordings of discussions. The dataset, obtained from EPFL's Entrepreneurship and Technology Commercialization lab, encompasses approximately 150 hours of dialogues for German entrepreneurs. Our aim is to construct a model to estimate the number of ideas generated by teams during their discussions. We start by training two models to extract emotions and MBTI personalities of speakers and then analyse and preprocess the entrepreuneurs' data. Subsequently, we implement various machine learning techniques to predict the quantity of novel ideas generated by each team.

The next Figure 1 illustrates the overall pipeline:

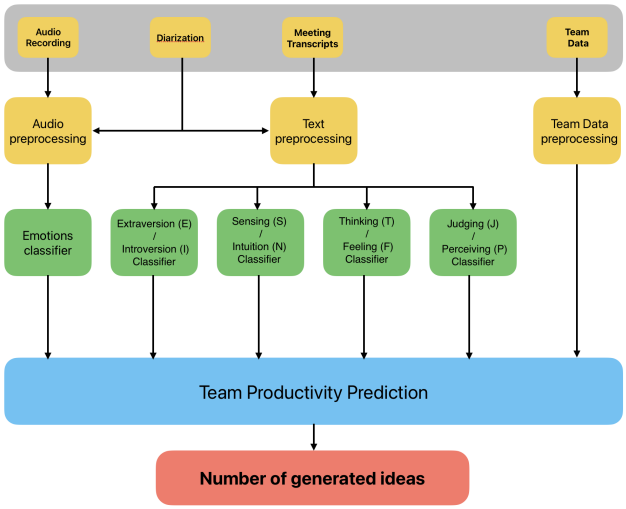

Fig. 1. Overall project approach

II. Features Extraction

A. Emotions recognition from audio recording

a) Audio preprocessing and features selection: To train our model, we use the Thorsten dataset. It's a dataset of 7 emotions recorded in german by Thorsten Muller based on his interpretation of what each emotion was. The Figure 2 illustrates the implemented pipeline to extract the emotions:

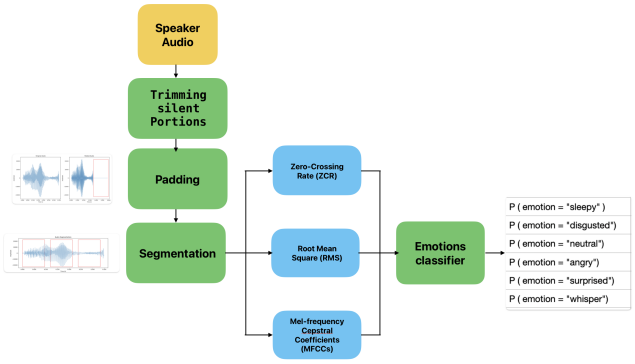

Fig. 2. Emotions detection pipeline

First, we start by trimming the silent portions of the audio (magnitude smaller than 20 db) in order to eliminate the noise, then we pad these audio files so that all samples have the same length to ensure uniformity. Subsequently, we apply segmentation with `FRAME_LENGTH = 2048`, representing the length of each audio frame, and `HOP_LENGTH = 512`, indicating the step size between consecutive frames. The sample rate (`sr`) was fixed at `8000 Hz`. It sets the frequency at which the audio signal is sampled, affecting the level of detail and information retained from the original audio recordings. The preprocessing was done using the `librosa` library in python. We used 32 features that can reconstruct a signal that is very close to the original [2], to predict the emotions:

- `Zero-Crossing Rate (ZCR)`: this feature, measures the rate at which the audio signal changes its sign. A higher ZCR may indicate more dynamic and rapidly changing audio content.
- `Root Mean Square (RMS) Feature`: this feature quantifies the energy of the audio signal. It represents

the magnitude of the signal and can provide insights into the overall loudness of the audio.

- `Mel-frequency Cepstral Coefficients (MFCCs) Feature`: these coefficients capture the spectral characteristics of the signal. They are widely used in speech and audio processing for representing the shape of the audio spectrum.

In the context of our emotion recognition model, these features contribute to capturing important aspects of the audio recordings, such as changes in signal intensity, energy distribution, and spectral characteristics.

b) Neural networks: The emotion detection task entails categorizing audio recordings into one of seven predefined emotions: "amused," "sleepy," "disgusted," "neutral," "angry," "surprised," and "whisper." Thus, the output layer utilizes a softmax activation, ensuring the model produces probability distributions across these emotion categories. We employed two Long Short-Term Memory **LSTM** [3] layers, providing the ability to capture temporal dependencies within the input audio sequences. The **LSTM** layers process the input data, followed by fully connected layers that refine the representations obtained from the first layers. Unlike traditional models using only fully connected architectures, this model is tailored to handle sequential data, making it well-suited for capturing temporal dependent patterns present in audio recordings. Since it is a mutli-class classification task, we used the Cross Entropy Loss and the RMSprop optimizer with a learning rate of 0.001.

c) Model performance evaluation: While our model performed admirably with an impressive 85% accuracy, it's crucial to consider its performance for all emotions illustrated by the Figure 3.

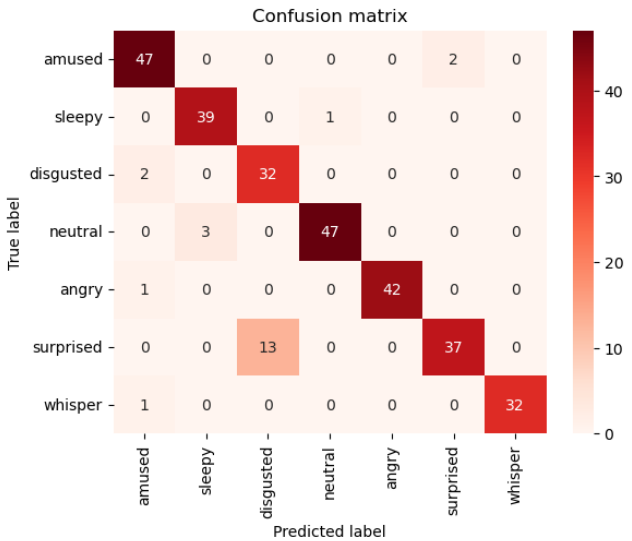

Fig. 3. Confusion matrix for emotions recognition

Notably, it consistently excelled in accurately predicting emotions like "whisper", "angry" and "sleepy". However, a more nuanced examination brings to occasional misclassifications around 2 to 3 samples per emotion for others emotions in the dataset. This confusion matrix demonstrates the robustness of our model. However, a significant portion of instances classified as "surprised" are instead identified as "disgusted," with 13 out of 50 samples affected.

d) Comments on the dataset: While we're thrilled about achieving high accuracy, it's crucial to highlight that this dataset reflects the beliefs of a single person regarding each emotion. Consequently, using the same model on real-world data, like entrepreneurs' talks, might yield less accuracy. The tone, especially for women, and the definition of each emotion can significantly differ from what Thorsten Muller recorded. This dataset was constructed by simulating emotions Throsten portrayed, but an alternative approach involves labeling genuine audio based on perceived emotions on movies for exemple. Both methods remain highly subjective. For potential datasets, a list is available. here, we can for example test the model on EmoDB or CMU-Multimodal SDK.

B. MBTI personalities using the transcripts

To train our model, we used the Twisty Dataset [4]. It's a dataset of 27 000 tweets in german and their associated MBTI labels. There are 16 MBTI categories based on 4 letters which are binary (Introverts (I) / Extroverts (E) , Sensing (S) / Intuition (N), Thinking (T) / Feeling (F) and Judging (J) / Perceiving (P)). The labeling of the data is based on the self-testing of the authors based on available MBTI personality tests on the internet like this one.

a) Text preprocessing and features selection: The text preprocessing steps involve loading and processing textual data for a machine learning model. The process includes:

- `Tokenization and Stopword Removal`: tokenize the text using the spaCy library for the German language, remove stopwords and extract lemmas to only capture meaningful information.
- `Sentiment Analysis`: utilize the VADER [5] sentiment analyzer to obtain sentiment scores (positive, negative, neutral, compound) for each text.
- `Word Encoding`: encode words by uniquely indexing every distinct word and pad sequences to ensure a fixed number of words.
- `Words Filtering`: filter out words based on their frequency, removing those occurring too frequently or infrequently and one-hot encode the selected words.
- `TF-IDF Vectorization`: use the TF-IDF vectorizer to convert the text into a sparse matrix. It assigns weights to words based on their frequency in a document and rarity across the entire corpus. The resulting matrix captures the importance of words in documents, representing documents as numerical vectors.
- `Dimensionality Reduction using LSA`: apply Latent Semantic Analysis LSA [6] using TruncatedSVD to reduce the dimensionality of the TF-IDF matrix.
- `Topic Modeling using LDA`: apply Latent Dirichlet Allocation LDA [7] for topic modeling, creating new features based on the topics identified.

These steps aim to prepare the text data for machine learning models, particularly for multi-output neural network models and non-neural network models. The applicability of each preprocessing step is dependent on the choice of the model being employed.

b) ML methods: Personalities of speakers can be deduced by employing a binary classification strategy for the four distinct MBTI personality pairings, or through a direct multi-class classification of the 16 unique MBTI labels. Initially, we opted for the multi-class approach using a BERT model pre-trained for this purpose. However, we later realized that since each MBTI personality pair operates independently, it was more effective to adopt various methods tailored specifically to deal with individual personality duos, without considering the interdependencies among different pairs.

- **K-nearest-neighbours:** deals with one personality duo (”Introvert” or ”Extrovert” for example) at a time, takes as input data features, to predict one of the 2 opposite personality types.
- **(Kernel) logistic regression:** takes as input text data features and predicts one of the personality duos at a time, just as with KNeighborsClassifier.
- **(Kernel) Support Vectors Machines:** similar to KNeighborsClassifier and LogisticRegression.
- **Neural-Networks:** use a Multi-Layer Perceptron model (SimpleMLP) to perform binary classification on each personality duo individually or use Multi-Task Multi-Layer Perceptron neural network model to perform multiple binary classification for different personality types at the same time.
- **Bert-pretrained model:** BERT [8] (Bidirectional Encoder Representations from Transformers) is a groundbreaking model in NLP introduced by Google. It takes data features converted into a format that BERT can understand and generates one output for each of the four personality duos.

It is important to mention that models predicting each of the four MBTI personality duos independently utilize multiprocessing for simultaneous computations, enhancing efficiency. This method involves creating separate processes for each duo, allowing parallel predictions and optimizing performance based on the specific modeling task at hand.

c) Models performance evaluation: The results can be summarized in this table, we tested our models on 5000 tweets from the dataset. We choose to implement all of these methods based on a previous work [9] and MBTI personnalities,

Methods	IE	SN	TF	JP	overall
Logistic regression	70%	82%	55%	62%	20%
K. Logistic regression	70%	82%	56%	63%	21%
Neural-network	70%	83%	56%	63%	20%
KNN	67%	83%	53%	61%	21%
SVM	72%	85%	64%	66%	30%
Bert	71%	82%	62%	64%	29%

TABLE I
BASIC MODELS ACCURACIES

d) Comments on the dataset: We got a lower accuracy on our overall prediction of the four personalities compared to emotions prediction from audio files. This can be justified, as the range of personalities is larger than just the 16 ones defined in the dataset. It is also very inaccurate to define a person as only being Extrovert or Introvert and not somewhere in between. This will motivate our choice of taking the probabilities of belonging to each class instead of the predicted labels themselves in our further work.

III. Predictions of ideas

The last step consists in using the trained models with the provided dataset of transcripts and audio from entrepreneurs' conversations to extraction the personalities and emotions of speakers. The Figure 4 describes all the implemented steps to train the final model to predict the team productivity:

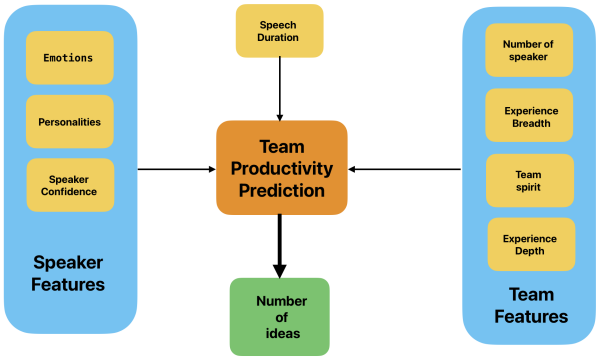

Fig. 4. Productivity prediction pipeline

Here are the features

- `Speaker features:`
  - Speaker personality : 4 probabilities representing introversion, intuition, thinking, and judging likelihood for each speaker.
  - Speaker emotion: 7 probabilities describing the distribution of emotions of each speaker.
  - Speaker Confidence: The percentage of confidence felt by a team member when presenting his ideas.
- `Team features:`
  - Team spirit: A binary indicator (0 or 1) reflecting the team's collaborative cohesion.
  - Experience breadth: The range of skills and knowledge diversity within the team.
  - Experience depth: The depth of expertise within the team members' individual domains.
  - Number of speakers: The count of speakers contributing to the team conversation.
- `Meeting feature:`
  - Speech duration: The number of minutes each meeting lasted.

In our final implementation, we adopt two distinct experimental strategies to forecast results using varied models. The initial experiment centers on predicting the mean number of ideas per individual speaker within his team. Conversely, the second experiment concentrates on the number of ideas generated collectively by teams. In each experiment, we apply two types of models: a kernel ridge regression model and several classification models. The kernel ridge regression is employed to establish a regression relationship between features and the predicted output. For classification, we utilize an array of models, including logistic regression, kernel logistic regression (K. logistic regression), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). The results from both sets of experiments are visualized and examined through box plots. Before showing the results, we define:

Model A: model trained on number of ideas per speaker.

Model B: model trained on number of ideas per team.

The Figures 5 and 6 illustrate the performance of 4 implemented classifiers for models A and B after running the cross validation. The tables II and III shows the result obtained after training all models

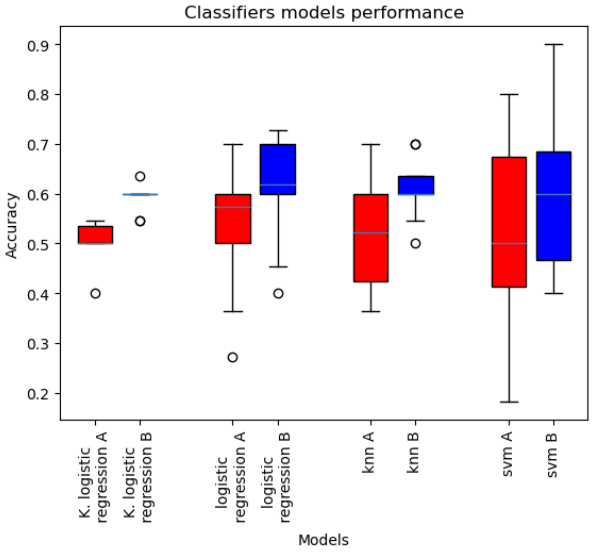

Fig. 5. Productivity prediction pipeline

Model	Accuracy			
	logistic regression	k. logistic regression	KNN	SVM
A	52%	58%	42%	67%
B	60%	62%	67%	58%

TABLE II
CLASSIFICATION MODEL

The Figure shows no significant differences in performance between classifiers, but when comparing models A and B, Model B (number of ideas per team) performs better, resulting in higher accuracy compared to Model A (number of ideas per speaker).

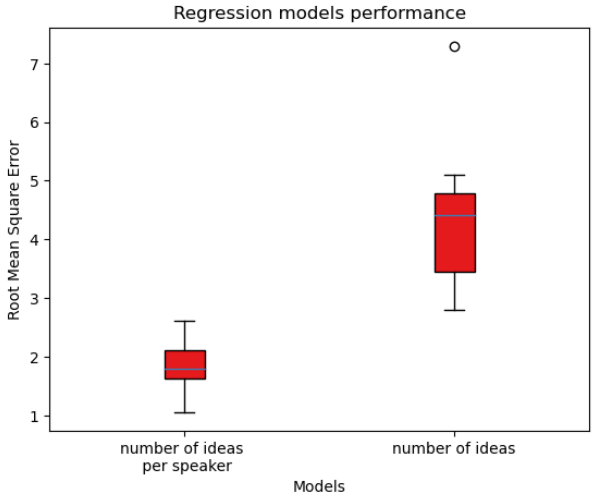

Fig. 6. Productivity prediction pipeline

Model	RMSE
A	1.7
B	4.4

TABLE III
REGRESSION MODEL

In Figure 6, we assess the performance of our regression models using Root Mean Square Error (RMSE) as the metric. Smaller RMSE values signify a closer model fit to the data. The plot presents a comparison of RMSE between models trained on two distinct outputs. It reveals that the model predicting the number of ideas per speaker exhibits higher accuracy indicated by the lower median RMSE.

IV. SUMMARY

In our project, we leverage machine learning techniques to assess the potential of entrepreneurial teams by predicting their personalities and emotions during discussions. Our approach involves working with multimodal data, including transcripts and audio recordings, in order to predict the number of generated ideas by teams. We achieved promising accuracy (85%) with our emotion recognition model. Meanwhile, our personality classifiers are showing varied outcomes. Subsequently, we combine these various features, along with team dynamics and meeting characteristics, to create a final model for predicting team productivity.

Throughout this project, we have encountered various challenges, which have provided valuable learning experiences. We gained expertise in processing both audio and text data. We have also learned how to implement recurrent neural networks like LSTM for emotion recognition, enhancing our understanding of sequential data analysis. We also learned how to choose and fine-tune the best models for our specific tasks. These valuable lessons have not only expanded our technical capabilities but also enriched our problem-solving and decision-making skills.

V. Ethical Risk

Examining entrepreneurs' personality traits and emotions raises complex ethical issues: including the risk of bias in decision-making processes and in the labeling of the data set because of stereotypes. An important aspect is the variability of personality traits and emotional expression across languages and its subjectivity. Such diversity could unintentionally lead venture capital funds to make decisions biased by entrepreneurs' nationality. For example, the attribution of specific traits to certain languages or nationalities could trigger predispositions in investors, influencing investment decisions and creating disparities between similar projects.

The stakeholders: The main ethical dilemma concerns potential intrusions into privacy and the formulation of psychological profiles based on sensitive data. The analysis of emotions and personalities may inadvertently classify individuals according to cultural stereotypes, due to the biases inherent in data labeling and the imbalances between categories in different countries.

To mitigate these potential effects: we have taken proactive measures in our model. First of all, the data has been anonymized to protect the privacy of the entrepreneurs. We balanced the data to avoid over-reliance on language-specific emotional traits or associations when training. In addition, once the training is done (on binary labels) we chose to make probabilities predictions such that our model is not wrongfully classifying people into categories.

The way emotion-based and personality-based data is labeled is inherently prone to bias, as it is always based on interpretation (by labelers if they label real data (audio or text) or by the questions which are asked for MBTI tests), and it's very difficult to avoid this due to the fact that there is no universal definition of what anger or joy is, for example.

To sum up everything, we have confidence in the potential of this model, to substantially enhance its accuracy while reducing biaises due to the subjectivity of datasets. To achieve improved performance, the primary emphasis should be placed on training with more appropriate datasets. For instance, concerning audio prediction, mitigating biases in dataset construction could be achieved by utilizing a combination of multiple datasets rather than a single one, thereby enhancing the model's training. We offer a compilation of datasets [10] that could be potentially leveraged in conjunction with our models (while retaining the same pipeline).

References

[1] N. Flammarion and M. Jaggi, "Cs433-machine learning course - epfl," 2023. [Online]. Available: https://github.com/epfml/ML_course.git

[2] S. Arzaghi, "Audio pre-processing for deep learning," 2020. [Online]. Available: https://www.researchgate.net/publication/347356900_Audio_Pre-Processing_For_Deep_Learning

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://www.bioinf.jku.at/publications/older/2604.pdf

[4] D. W. . P. B. Verhoeven, B., "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling," 2016. [Online]. Available: https://www.uantwerpen.be/en/research-groups/clips/research/datasets/

[5] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[6] P. Kherwa and P. Bansal, "Latent semantic analysis: An approach to understand semantic of text," *Journal Name*, vol. Volume Number, no. Issue Number, p. Page Range, Year of Publication, please fill in the missing details such as Year, Journal Name, Volume, Issue, Page Range, and DOI if available. [Online]. Available: URLtotheArticle

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003, accessed: Date of Access. [Online]. Available: URLtotheArticle

[8] J. D. M.-W. C. K. L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf

[9] P. M. K. Prajwal Kaushal, Nithin Bharadwaj and A. K. Koundinya, "Myers-briggs personality prediction and sentiment analysis of twitter using machine learning classifiers and bert," 2021. [Online]. Available: https://www.mecs-press.org/ijitcs/ijitcs-v13-n6/IJITCS-V13-N6-4.pdf

[10] N. Barazida, "40 open-source audio datasets for ml," 2021. [Online]. Available: https://towardsdatascience.com/40-open-source-audio-datasets-for-ml-59dc39d48f06