Podcasts Are Not All You Need 🚀

Tristan Jacobs
jacobst@informatik.uni-marburg.de

Bashir Hussein
husseinb@informatik.uni-marburg.de

Christian Kujath
kujath@informatik.uni-marburg.de

Ryan Ford
ford@informatik.uni-marburg.de

Abstract

Cryptocurrencies are digital assets acting as a medium of exchange (MoE) which are known for high volatility and (community-driven) hype cycles. In an attempt to track this volatility, we conduct an exploratory sentiment analysis on audio data - podcasts. Finally, we plot this sentiment data against cryptocurrency prices as a basic price prediction visualization.

1 Introduction

Many trends have emerged over the past year, partially due to the COVID-19 pandemic. For one, financial market turbulences sunk and lifted asset prices to historical lows and all-time-highs (Lin). Cryptocurrencies in particular, a technologically-driven category of financial assets, have largely benefited from this unexpected volatility. Most prominently, Bitcoin (BTC), Ethereum (ETH), Dogecoin (DOGE), and derivatives of these currencies have been affected. These trends have not yet been able to be adequately foreseen using automated, social media-based sentiment analysis (e.g. Vaneck Social Sentiment BUZZ ETF) (BUZ). We believe this could be the result of the inherent properties of such data: social media data is often too noisy and unreliable, due to the promotion and emergence of trends through virility and the platform's algorithms, rather than more sound, higher quality analysis (Abdullah et al., 2019). Podcasting has been another trend, gaining increasing popularity (Pod). Since podcasting has low technological entry barriers, albeit higher than those of social media forums, there has been a surge of available podcasts. Because podcasts are typically entirely free to use (for consumers), the competition for user's attention has increased significantly. We believe this has resulted in the quality of content being more easily identifiable through simpler rating/popularity mechanisms.

We postulated that these factors could render podcasts a great source for sentiment analysis: Podcasts are freely available on the internet, have high quality advantages over regular social media-based data, and offer new technological challenges to tackle. We were interested in seeing to what extent we can extract this perceived higher quality from the data, which could perhaps be used in time series models in future work to predict macro price changes of the affected assets over time.

2 Methods

2.1 Overview

Figure 1 provides an overview of our data pipeline. This pipeline is implemented in the `get_sentiments` function in `CryptoSentimentAnalysis.py`. It takes a collection of URLs as input and starts by downloading the video files, which are converted to audio files after downloading. In the next step, each audio file is cut into clips of equal length. Our trained Wav2Vec model will then transcribe each clip. Based on this text we assign a coin label to the clip. Then, we use a software called Praat (Boersma and Weenink, 2018) to extract certain audio features from each clip. Finally, we predict the sentiment of a clip based on the previously generated text and extracted audio features. Some of the steps are described in more detail below.

2.2 Data Collection

As mentioned above, our pipeline uses a Wav2Vec model for the transcription and a sentiment analysis model for the sentiment prediction. Both models were trained by us with

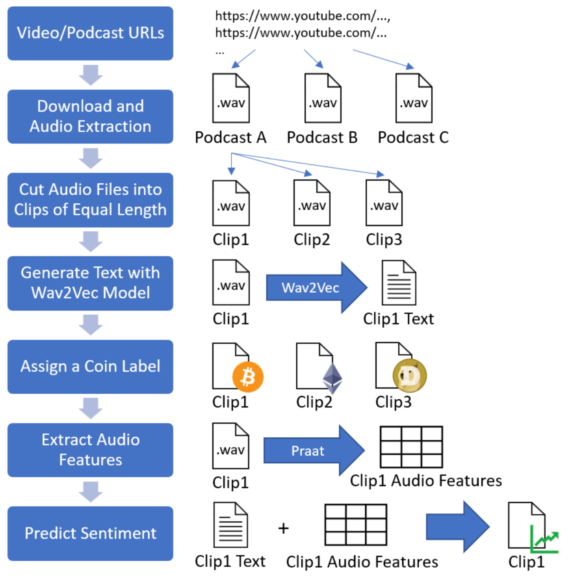

Figure 1: Our data pipeline: from video URLs to sentiments.

data curated by ourselves. Data collection, processing, and labelling were by far the most time consuming parts of our project. Because no data sets currently exist containing transcription data of podcasts in the cryptocurrency domain, our team collected and manually labelled data for over 1500 audio clips, including transcription, topic, and sentiment.

Due to initial assumptions about our available computing power, we split each podcast into 12-15 second clips containing 30 words each. The initial transcription data, and most importantly, the time stamps, came from YouTube's automatic transcriptions. Figure 2 shows the data acquisition process in detail. When splitting YouTube's subtitle text into chunks of 30 words, we also save the start and end timestamps for each chunk. This helps when we manually review and correct the YouTube transcription and is also necessary for the audio feature extraction. After the YouTube transcription is corrected, audio files are cut into clips based on chunk time stamps and the Wav2Vec model is trained. Output of the Wav2Vec model is later used as training input for the sentiment analysis model in addition to audio features, also extracted based on time stamps.

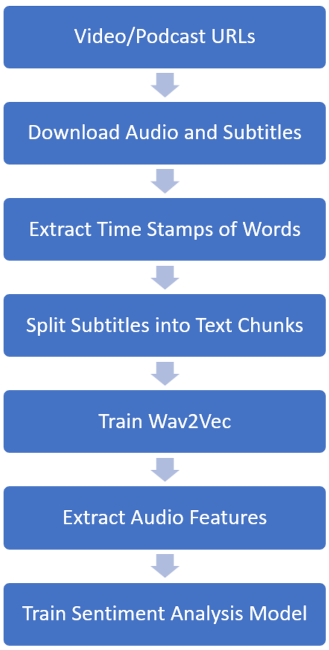
Video/Podcast URLs
Download Audio and Subtitles
Extract Time Stamps of Words
Split Subtitles into Text Chunks
Train Wav2Vec
Extract Audio Features
Train Sentiment Analysis Model

Figure 2: Pipeline for data collection

2.3 Wav2Vec2 Fine-tuning

Wav2Vec2 as presented in (Baevski, 2020) was the architecture we chose to train our Speech-To-Text model on. Wav2Vec2's architecture learns how to create meaningful audio representations for Connectionist Temporal Classification (CTC). This means that by harnessing a pretrained instance of Wav2Vec, we can fine tune a single fully-connected layer on it's outputs to learn how to model new words. As state-of-the-art speech to text is achieved with Facebook's pretrained instance (`facebook/wav2vec2-large-960h`), we used this to train our model. Without fine tuning, the model achieved a WER of about 27% on our corrected transcriptions. After training[1] on our data points (30/10 split), we were able to arrive at a WER of 13.1%.

2.4 Sentiment Analysis Pipeline

2.4.1 Transcript Generation

After downloading the podcasts, they are automatically split up into clips of 15 seconds each. These audio clips are then fed to our Wav2Vec2 Model to generate the transcription for every audio clip.

[1] `scripts/Wav2Vec_Train.ipynb|`

2.4.2 Coin Prediction

Once the transcription is generated, a simple regex is performed to determine the coin of interest. As discussed in section 4, this regex search for coin could be replaced with a HMM to track the topic with time dependencies.

2.4.3 Audio Features

We extracted Audio Features (Pitch-05-Quantile, Pitch Range, Jitter ..etc) using Praat, which are used in conjunction with the transcription for our sentiment analysis model. Those features were similarly used and recommended by (Mairano et al., 2019).

2.4.4 Sentiment Analysis

Our final sentiment model first transformed our text into a Tf-Idf representation concatenated with our audio features, before being trained on a Multilayer Perceptron Classifier with 100 hidden layers and the ReLU activation function. We attempted to use different classification models but assume that a finer grained textual representation will be able to deliver better results in the future instead. In the end, our sentiment model had an accuracy of 56%. The confusion matrix is depicted in figure 3.

We believe a large issue facing our performance was the labelling of our dataset. As we had to manually label our data, our time was severely limited and we could not explore better textual representations and sentiment models. In particular, we would have change some things about our labelling process in retrospect. We started with a text window which was too small, making the sentiment difficult to read even for humans. This was compounded by the fact that the data is from dialogue, making it contain many filler words that further works could try to explicitly remove.

3 Evaluation and Results

We evaluate our sentiment analysis process pipeline's accuracy by comparing achieved results, i.e. sentiment labels per coin, date, and clip, to the actual price development. As source of truth, we use price data from Yahoo Finance derived using

Due to a podcasts' property of being consumed any time, sentiments presented are - unlike many commonly used data sources for SA, e.g. news tickers, - not necessarily focused on the short-term, but mid or even long time horizons. Thus, we are testing both short- and mid-term performance in separated tests. Since cryptocurrency markets are moving fast, changing rapidly (due to political interventions (Chi), hype cycles, or technological changes like forking), we decided to determine the mid-term time range as one week.

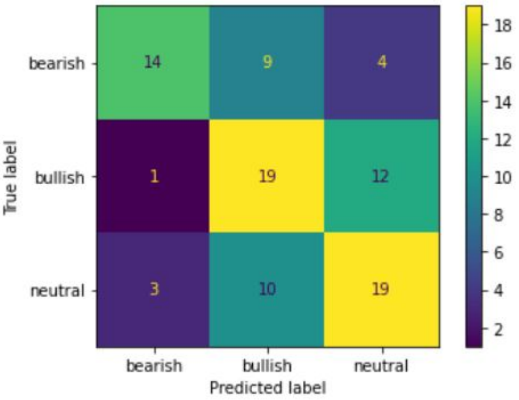

Figure 3: Confusion Matrix.

Coin	Correct	Incorrect	Ratio
BTC	2	6	0.3
DOGE	4	4	0.5
ETH	6	2	0.75
Ratio	12	12	0.5

We have achieved the following results for interpreting podcast sentiments in the short-term (i.e. day-basis):

Furthermore, results of generated coin sentiments on the mid-term (i.e. week-basis):

4 Discussion and Future Work

Our results are highly dependent on the efforts and quality of work put into the various underlying datasets, models, and techniques. With regards to the Wav2Vec model, we can confirm that good results are quickly achievable. 13.1% WER is very competitive, especially considering the nature of the data - most CTC datasets use clear speech data, such as that of audio books. Achieving a lower WER would require proportionally extra time for manual labelling.

Coin	Correct	Incorrect	Ratio
BTC	1	4	0.2
DOGE	1	4	0.2
ETH	0	5	0.0
Ratio	2	13	0.133

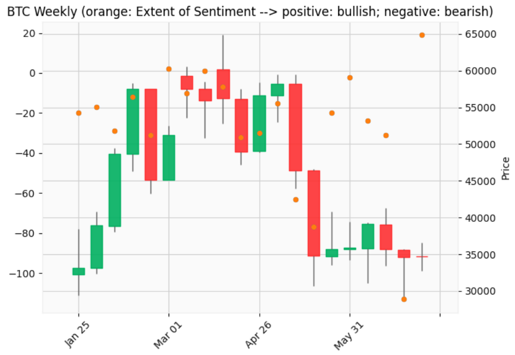

Figure 4: Sentiment Prediction Results for BTC (as orange dots, right y-axis) and Observed Prices (candle sticks; left y-axis) over Time.

Furthermore, deeper understanding of the multiple models underlying our sentiment analysis model is required - especially with regard to parameter optimization. Our original assumption that podcasts are of higher quality than ”usual” sentiment analysis based on social media data is questionable.

5 Conclusion

We have presented an exploratory pipeline for sentiment analysis on audio data, applicable on two vastly growing trends of podcasting and democratization of cryptocurrency investment. Our sentiment analysis included various technological challenges, which we tackled by creating custom training datasets (for Wav2Vec), extracting audio features (using Praat), and applying a sentiment analysis model. As our results show, extracting sentiment from dialogue is a hard task, and extrapolating it to currency price currencies is even more difficult.

We believe that there are two obvious improvements to be made. One of these would be to not evaluate every clip independently, but rather in sequence (for example with a hidden markov model or RNN). Next, we would train word embeddings on our text and experiment with different textual representations for the sentiment model. Using a Tf-Idf representation turned out to be extremely restrictive as a lot of the sentiments are captured in comparisons, where the order of words in the document matters.

Further work in the field of language transcription, audio feature extraction, and in-depth model parameter optimization for sentiment analysis are needed.

References

As podcasts continue to grow in popularity, ad dollars follow.

Bitcoin sinks to two-week low as china intensifies crypto mining crackdown.

Buzz, the etf of social-media darlings, drops in trading debut.

Crypto investors ’should be prepared to lose all their money,’ top uk regulator warns.

Nor Aniza Abdullah, Ali Feizollah, Ainin Sulaiman, and Nor Badrul Anuar. 2019. Challenges and recommended solutions in multi-source and multi-domain sentiment analysis. *IEEE Access*, 7:144957–144971.

Mohamed Auli Baevski, Zhou. 2020. A framework for self-supervised learning of speech representations. *Computation and Language*, 6.

Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer [computer program]. version 6.0. 37. *Retrieved February*, 3:2018.

Paolo Mairano, Enrico Zovato, and Vito Quinci. 2019. Do sentiment analysis scores correlate with acoustic features of emotional speech? In *AISV Conf.*