Pipeline Data processing

1. Dataset gathering
   1. Tweets containing the word Bitcoin 2017-2019
   2. url: <https://www.kaggle.com/franciscoarchilla/tweets-containing-the-word-bitcoin-20172019>
   3. cut to 8 Mart 2018 🡪 4 November 2018
   4. every month contains 700K to 1Million tweets
   5. all the gaps that occurred in the dataset have been filled and checked if they were correct
2. Language Labeling
   1. All tweets languages are labeled using the library whatthelang
      1. Source https://github.com/indix/whatthelang
   2. Papers that use English tweets :
      1. CNN-based multivariate data analysis for bitcoin trend prediction
      2. Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data
   3. Vader sentiment analyzer and BERTtweet is learned by using the English language -> we use English only text
      1. Paper: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text
      2. BERTweet: A pre-trained language model for English Tweets
   4. Optional: do a language analysis of the tweets you collected
3. Vader Sentiment analysis
   1. Preprocessing :
      1. Normalizing text to utf-8 format so it can be processed by vader
      2. Removing contractions
      3. Lower case
      4. Removing www and http://
      5. Removing @username
      6. Removing whitespaces
      7. Removing #
      8. Trimming \
      9. Replacing punctuations
      10. Removing digits
      11. Source : <https://towardsdatascience.com/are-you-scared-vader-understanding-how-nlp-pre-processing-impacts-vader-scoring-4f4edadbc91d>
      12. Source: A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19
   2. Calculating vader score
4. Checking if tweet is generated by a bot
   1. Using technique described in “Predicting Bitcoin price fluctuation with Twitter sentiment analysis” p18
5. BERT tagging
   1. Tweet is not preprocessed since BERTweet has its own normalizer
   2. BERT base is used (time constraints)
   3. Text needs to be english
   4. SOURCE/ BERTweet: A pre-trained language model for English Tweets

Research questions:

1. What is the impact of weekdays in model?
2. Impact of different time intervals? Is there a Timezone effect and does this disappear when the interval is larger than 1day?
3. Vader Compound score vs individual scores?
4. BERT is used to calculate tweet sentiment, that then is used for price prediction,
   1. This paper indicates poor performance:
      1. The reason for such poor performance is BERT is pre-trained using Wikipedia articles, which has well-written articles, i.e., which follows grammar rules. In comparison, the tweets are usually written using locally spoken slang like ‘‘wanna’’, ‘‘gonna’’, ‘‘YOLO’’, which generally doesn’t follow strong grammar rules making the classification problem harder.
      2. They used it for sentiment analysis and did not use the whole vector (check this)
   2. My question How will BERTTweet perform?