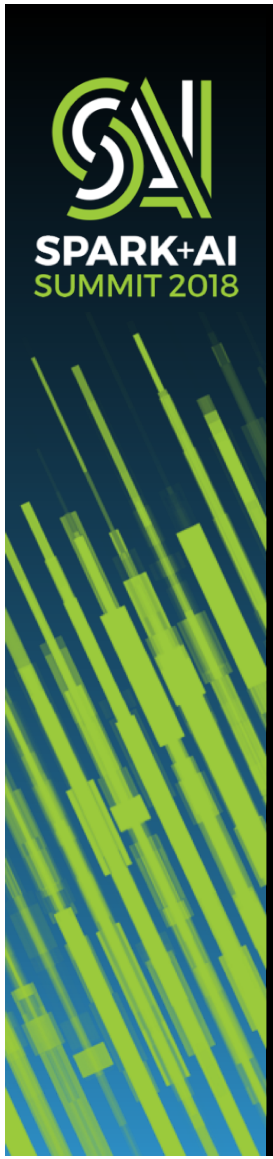


Deep Learning based Opinion Mining for Digital Currency Forecasting

Joyesh Mishra
Shibani Singh
Intel Corporation

#AISAIS15

Agenda

- Current Trends
- Workflow overview
- Execution & Dataflow
- Model Evaluation & Learnings
- Further research
- Key Takeaways

What this talk is not about!

- It is not going to make you super rich.
- It is not going to teach you about picking stocks or timing the market.
- It is not making any predictions whatsoever.

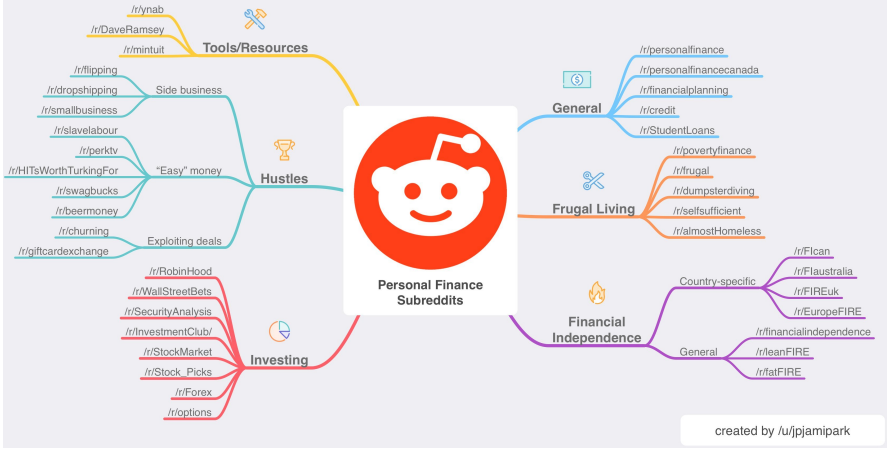
This talk is primarily for showcasing practical applications in Sentiment Analysis & NLP using Deep Learning Techniques.

Current Trends & Platforms

- Sentiment Analysis moves organizations, businesses, people and countries
- Maturity of large scale machine learning and democratization of deep learning techniques
- State of AI in Sentiment Analysis

Dataset

- Data Collection
- Subreddit selection & Twitter hash tags filters



Sources:

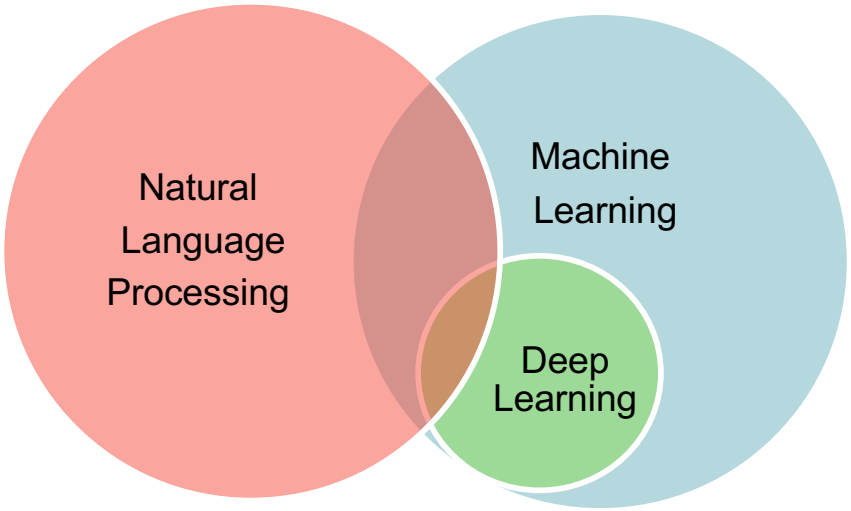
1 https://www.reddit.com/r/coolguides/comments/8fnj6g/financial_subreddits_guide/

2 https://ritetag.com/best-hashtags-for/investing

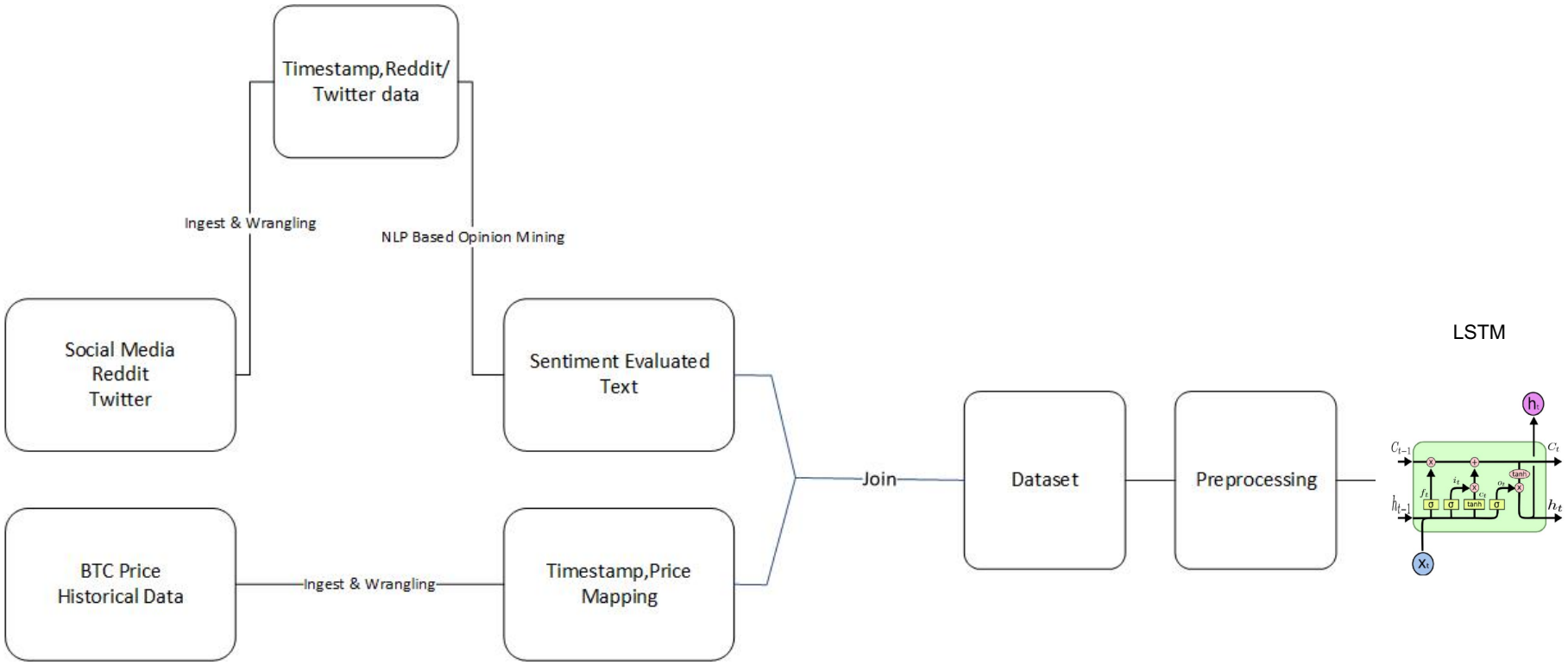
Workflow Overview

Natural Language Processing

- Statistical & rule based techniques
- Deep Learning (RNN, LSTM, GRU)



Workflow Overview

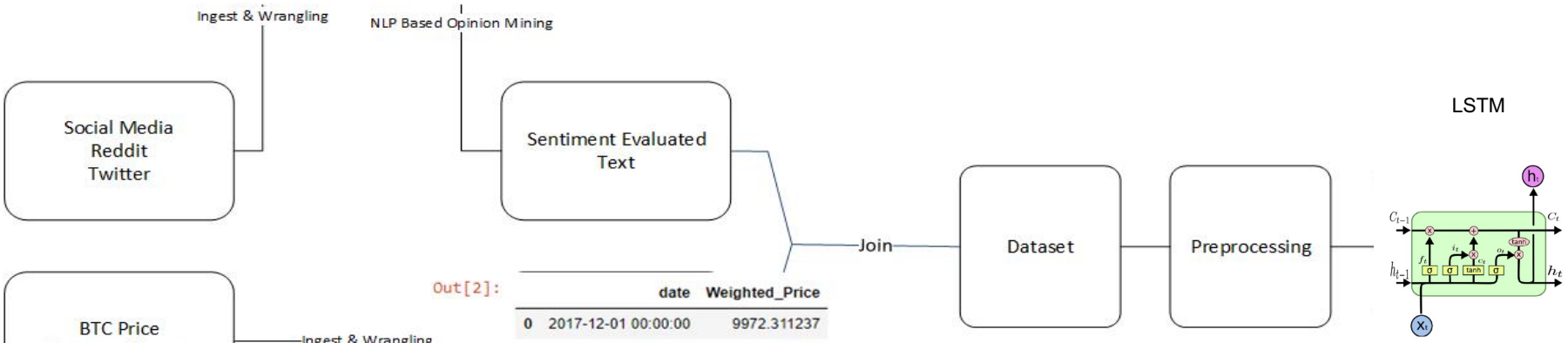


Workflow Overview – Data Acquisition

```
all_bitcoin_subreddit.tail()
```

Out[3]:

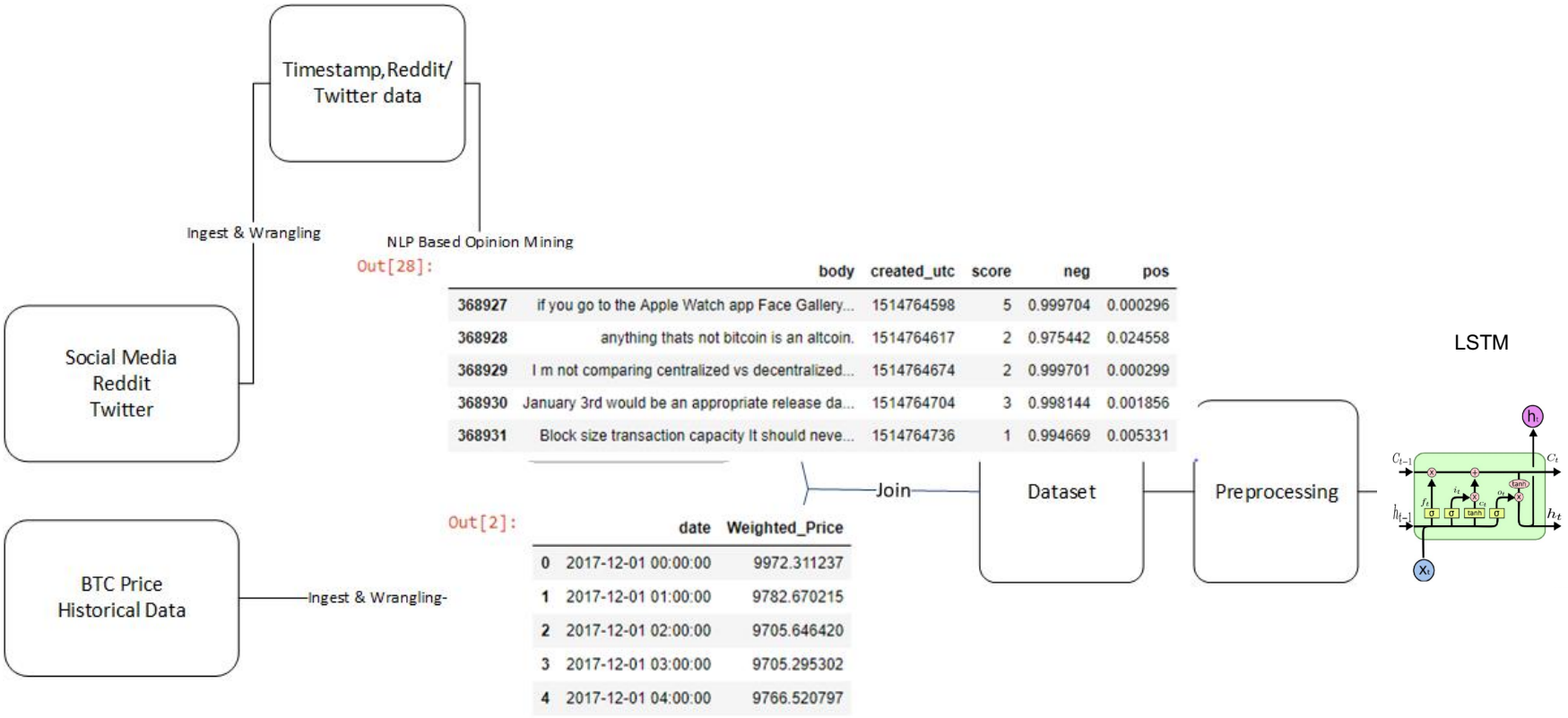
	body	created_utc	score
368927	if you go to the Apple Watch app Face Gallery...	1514764598	5
368928	anything thats not bitcoin is an altcoin.	1514764617	2
368929	I m not comparing centralized vs decentralized...	1514764674	2
368930	January 3rd would be an appropriate release da...	1514764704	3
368931	Block size transaction capacity It should neve...	1514764736	1



Out[2]:

	date	Weighted_Price
0	2017-12-01 00:00:00	9972.311237
1	2017-12-01 01:00:00	9782.670215
2	2017-12-01 02:00:00	9705.646420
3	2017-12-01 03:00:00	9705.295302
4	2017-12-01 04:00:00	9766.520797

Workflow Overview – NLP



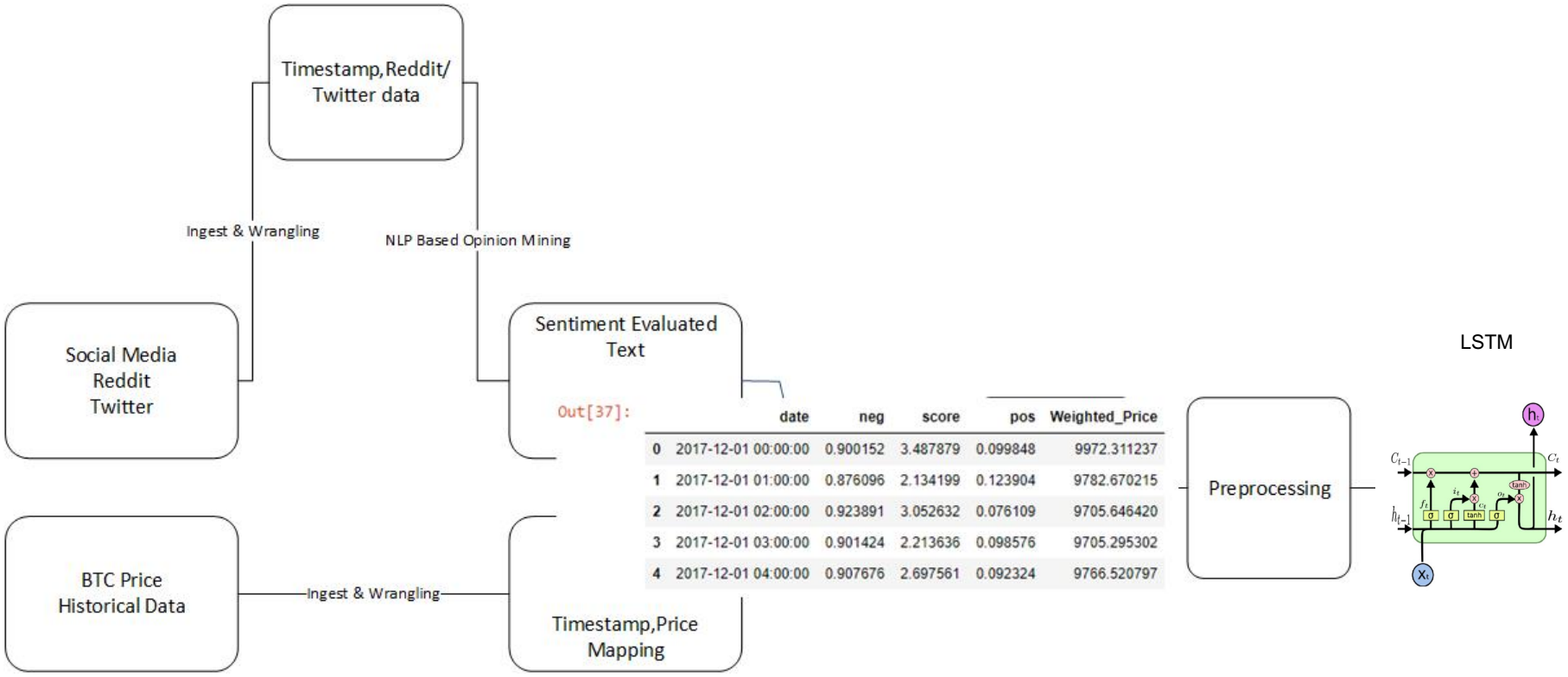
Out[28]:

	body	created_utc	score	neg	pos
368927	if you go to the Apple Watch app Face Gallery...	1514764598	5	0.999704	0.000296
368928	anything thats not bitcoin is an altcoin.	1514764617	2	0.975442	0.024558
368929	I m not comparing centralized vs decentralized...	1514764674	2	0.999701	0.000299
368930	January 3rd would be an appropriate release da...	1514764704	3	0.998144	0.001856
368931	Block size transaction capacity It should neve...	1514764736	1	0.994669	0.005331

Out[2]:

	date	Weighted_Price
0	2017-12-01 00:00:00	9972.311237
1	2017-12-01 01:00:00	9782.670215
2	2017-12-01 02:00:00	9705.646420
3	2017-12-01 03:00:00	9705.295302
4	2017-12-01 04:00:00	9766.520797

Workflow Overview – Merged Dataset



	date	neg	score	pos	Weighted_Price
0	2017-12-01 00:00:00	0.900152	3.487879	0.099848	9972.311237
1	2017-12-01 01:00:00	0.876096	2.134199	0.123904	9782.670215
2	2017-12-01 02:00:00	0.923891	3.052632	0.076109	9705.646420
3	2017-12-01 03:00:00	0.901424	2.213636	0.098576	9705.295302
4	2017-12-01 04:00:00	0.907676	2.697561	0.092324	9766.520797

Workflow Overview – Sentiment Modeling

NLP – Approach 1

- Vader[1]: "lexicon and rule-based sentiment analysis tool that is *specifically attuned to sentiments expressed in social media"*
- Available with NLTK

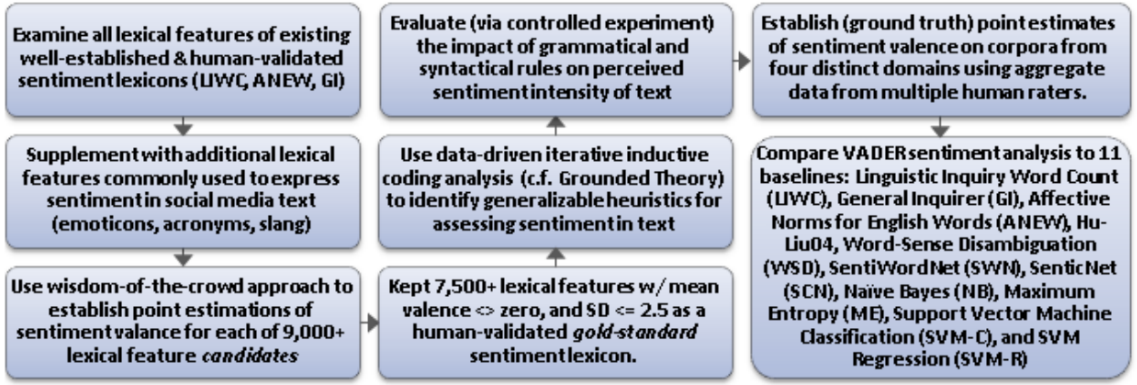


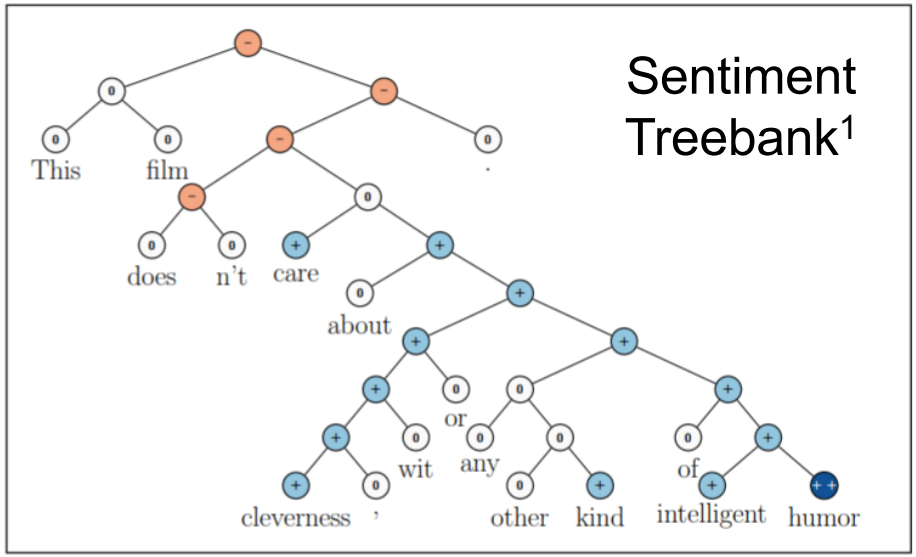
Figure 1: Methods and process approach overview.

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Workflow Overview – Sentiment Modeling

NLP – Approach 2

- Stanford CoreNLP
- Deep Recursive Model(RNTN) trained on Sentiment Treebank



Recursive Neural Tensor Networks have a tree structure with a neural net at each node

[1] Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts Stanford University, Stanford, CA 94305, USA

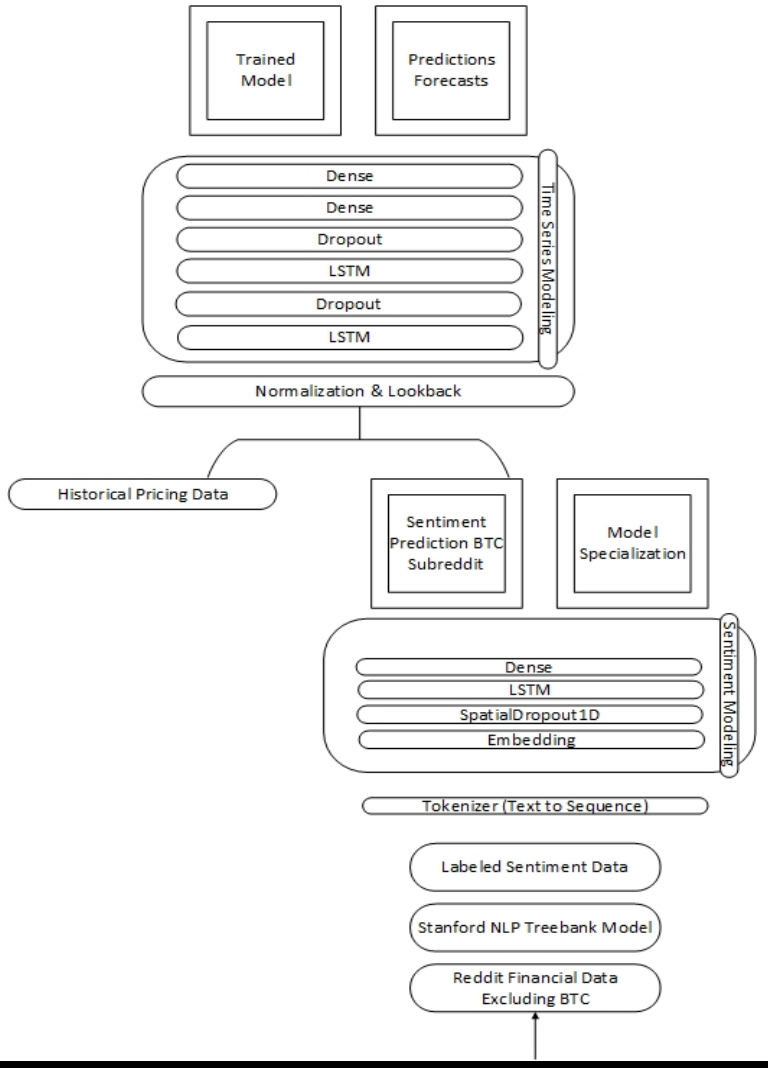
Workflow Overview – Sentiment Modeling

NLP – Approach 2 (Continued)

- Label all data for financial Subreddit comments & tweets (excluding BTC) using Stanford NLP Treebank
- Train an LSTM on the data above
- Predict sentiments – using Trained LSTM Model

Complete Pipeline

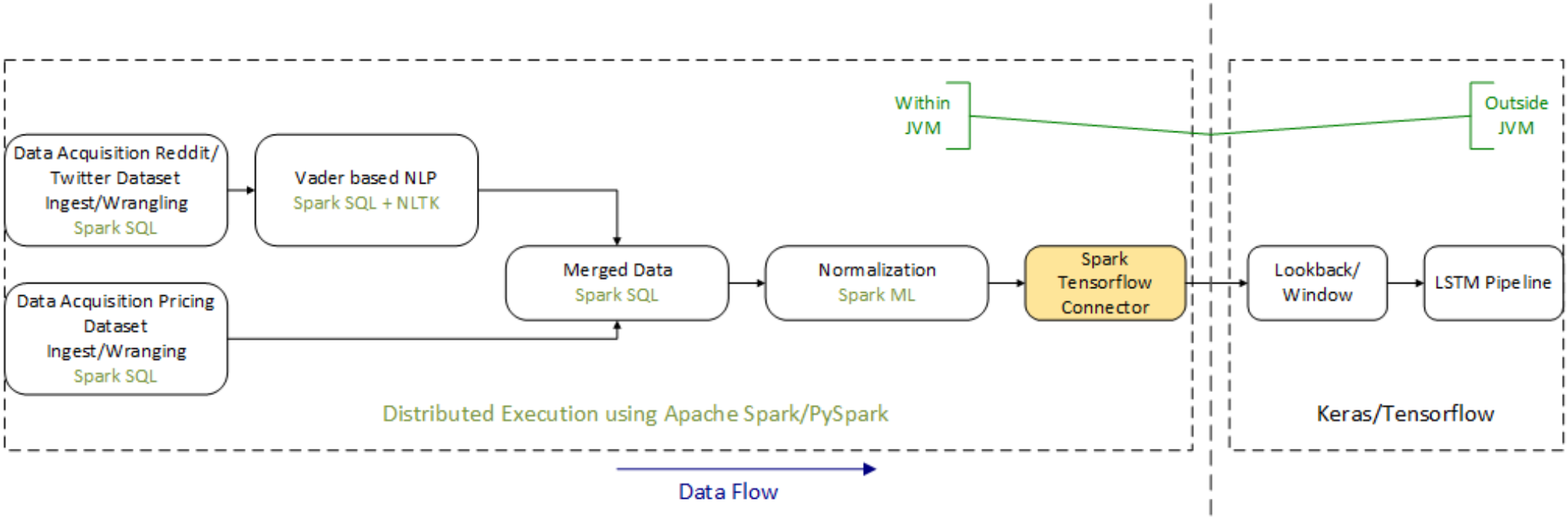
- Embedding + LSTM for Sentiment Modeling
- LSTM with Recurrent Dropout for Price Prediction and Time Series Modeling[1]



[1] A Theoretically Grounded Application of Dropout in Recurrent Neural Networks Yarin Gal, Zoubin Ghahramani arXiv:1512.05287v5 [stat.ML]

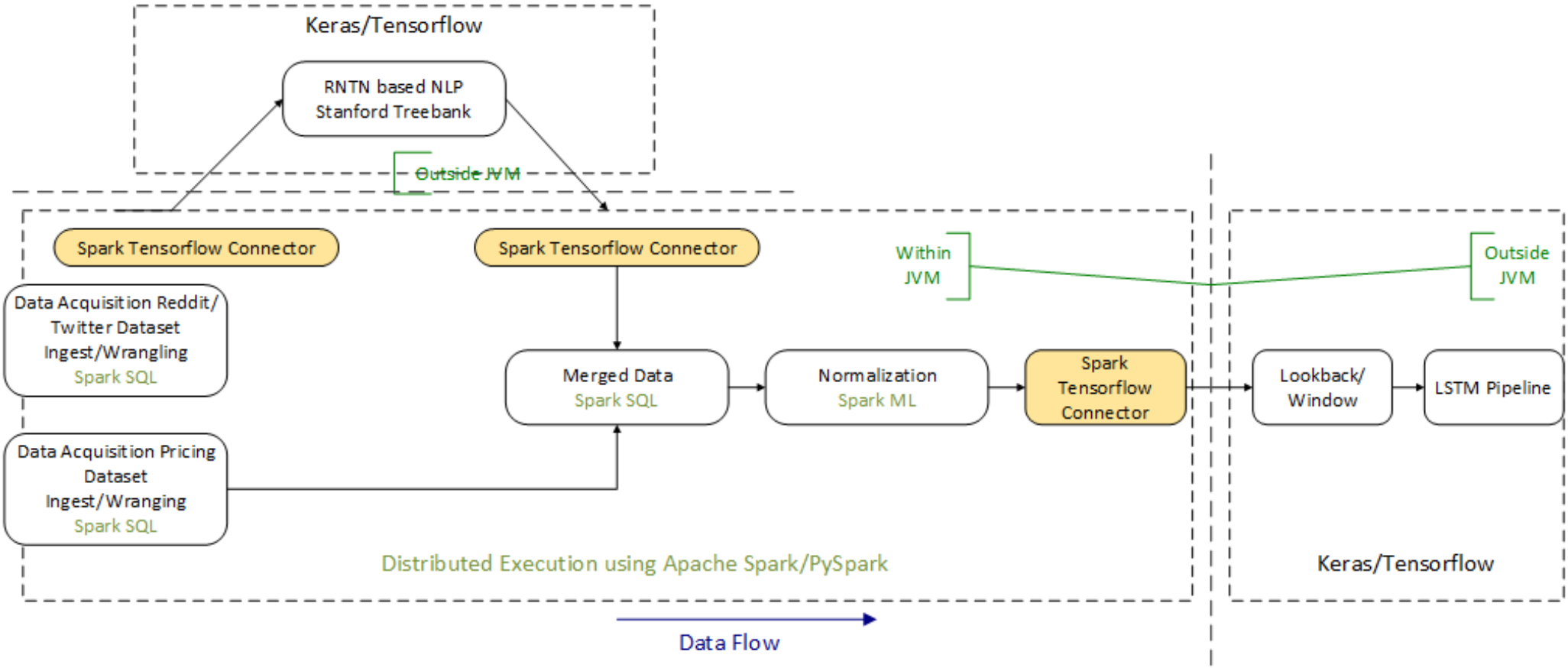
Execution

Iteration 1 - Spark, Keras (TF) & Spark-Tensorflow-Connector with Vader for Sentiment Analysis



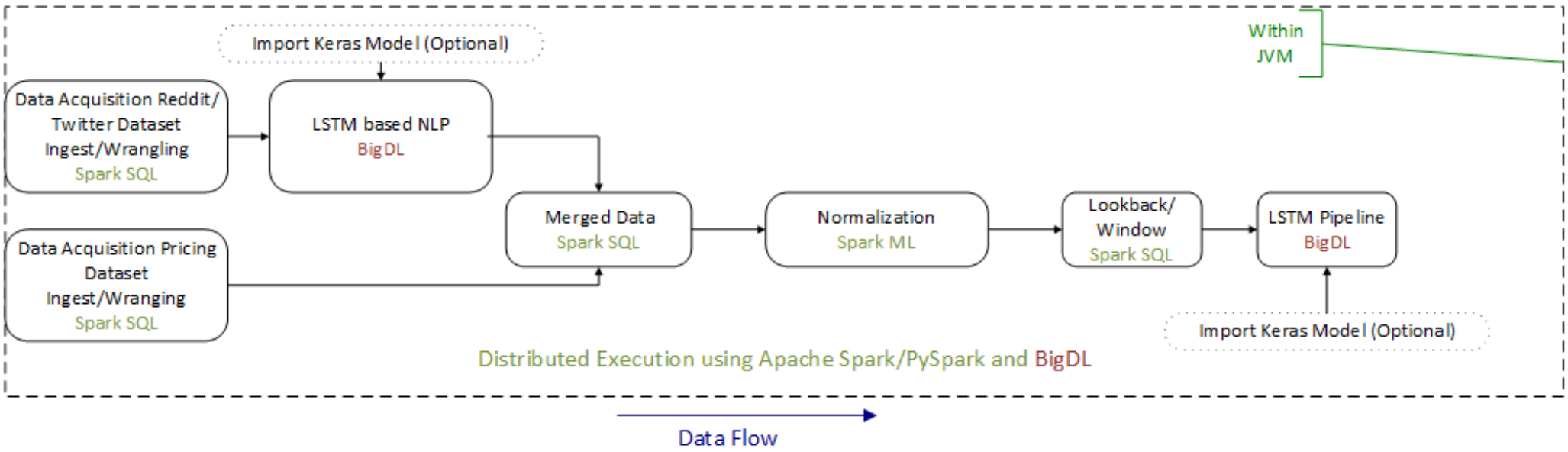
Execution

Iteration 2 - Spark, Keras (TF) & Spark-Tensorflow-Connector with Keras(TF/LSTM/Stanford Treebank) for Sentiment Analysis



Execution

Combined Workflow with BigDL + Keras Style APIs using Spark (Proposed Work in Progress)



- BigDL is a distributed deep learning library for Apache Spark
- Efficient Scale out leveraging Spark and Spark ecosystem, Synchronous SGD, All-Reduce
- Rich DL API/Layers Support (Native). Ability to re-use Keras Models (via Export/Import)
- Provides Keras Style APIs in Python/Scala for users familiar with Keras (Based on Keras 1.2.2)

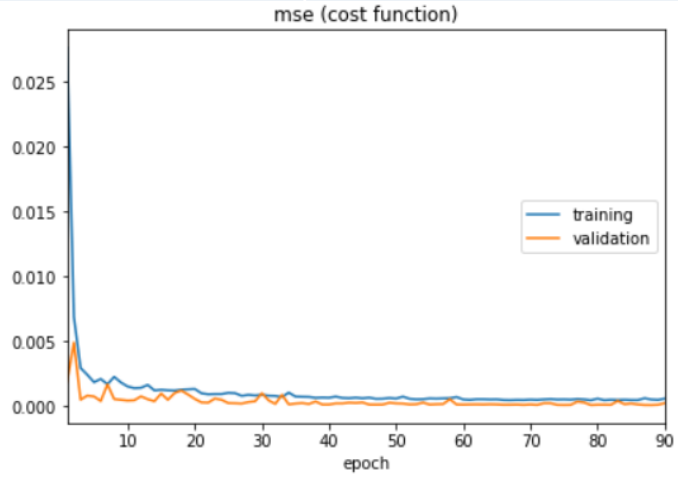
Model Evaluation & Learnings

- Adding transaction volume per time interval improves the accuracy
- Normalize all features to reduce any feature importance bias
- Different window lengths could be used to experiment multiple models (Best observed look back window ~ 22 hours)
- Further fine tuning of the layers could improvise predictions.

Model Evaluation & Learnings

LSTM based Sentiment + LSTM (value prediction)		Rule Based Sentiment (Vader) + LSTM (value prediction)	
Mean Absolute Error (Normalized data)	0.0084	Mean Absolute Error (Normalized data)	0.0110
Loss	0.0001	Loss	0.0002

The workflow inclusive of deep learning based sentiment prediction slightly outperforms the one where a rule based approach is followed for sentiment evaluation.



Further Research in Sentiment Analysis using LSTM and RecNN

- **Document Level Sentiment Classification**
  - Words Embedding → Dense Document Vectors → LSTM
  - Use Attention Mechanism and Non-Neural Classifiers (SVM)
- **Sentence Level Sentiment Classification**
  - Subjectivity Classification
  - RNTN, TG-RNN, TE-RNN, DCNN, CharSCNN
- **Aspect Level Sentiment Classification**
  - Aspect Extraction, Entity Extraction
  - AdaRNN, TD-LSTM/TC-LSTM
- **Emotional Analysis, Sarcasm Detection, Multi-lingual Sentiment Analysis**
- **Multi-Modal (Combining Textual, Visual, Acoustic etc.)**

Citation: Deep Learning for Sentiment Analysis : A Survey, 2018, Lei Zhang, Shuai Wang, Bing Liu, http://arxiv.org/abs/1801.07883

Key Takeaways

- Deep Learning based NLP techniques are comparable to statistical and traditional models
- Iterated training and model tuning could be achieved on large datasets due to maturity of DL frameworks being able to leverage each other (Keras, Tensorflow, Spark, BigDL)
- Model specialization and domain based training is really helpful in improving learning

Libraries

- Spark Tensorflow Connector
  - https://github.com/tensorflow/ecosystem/tree/master/spark/spark-tensorflow-connector
- BigDL
  - https://github.com/intel-analytics/BigDL

Thank You!