Algorithmic Trading using KAP and Twitter Sentiments with Machine Learning

Feyza Nur Bulgurcu 150117033 feyzabulgurcu@marun.edu.tr
Fatih Emin Öge 150118034 fatihoge@marun.edu.tr
Ahmet Akıl 150118038 ahmetakil@marun.edu.tr

Big Data and Text Analytics Research Lab. (@BIGDaTA_Lab)

Assoc. Prof. Dr. Murat Can GANİZ

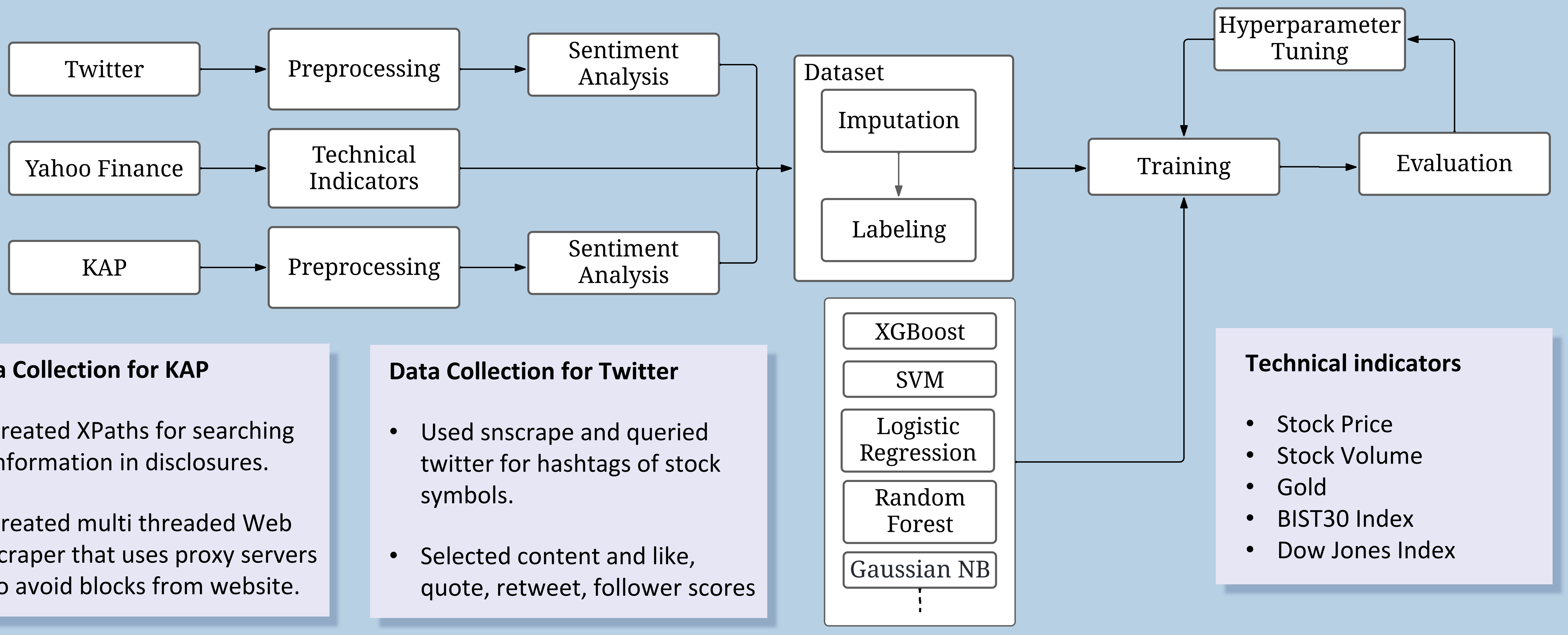
INTRODUCTION

- Algorithmic Trading can be defined as a trading method with pre-programmed trading instructions. It is an automated system that uses several metrics such as price, volume, season (time) data.
- Using the stock price alone is not healthy for future predictions. Thus, we plan to use other data sources as well as stock price, to see if the profits can be increased even further.

PROPOSED SOLUTION

- Machine learning models to predict direction of stocks in BIST30 index.
- Trained by using financial price data, financial sentiment analysis from Twitter and official KAP news (public disclosure platform for BIST).
- Calculating different time intervals for price change to compare the models' effectiveness over short-term or long-term.

CONTROL FLOW



Labelling

- Classification of data points with respect to the daily change
- Used 2 and 6 classes for different experiments

Data Collection for KAP

- Created XPaths for searching information in disclosures.
- Created multi threaded Web scraper that uses proxy servers to avoid blocks from website.

Data Collection for Twitter

- Used snscrape and queried twitter for hashtags of stock symbols.
- Selected content and like, quote, retweet, follower scores

Technical indicators

- Stock Price
- Stock Volume
- Gold
- BIST30 Index
- Dow Jones Index

Preprocessing

- Make Lowercase
- Remove punctuations and digits
- Remove URLs
- Remove stop words

Data Collection

- Fetched all tweets[1][2] containing hashtags of symbols of BIST30 stocks
- Fetched all KAP[1] news for the interval 2016-2021 and prepared them for sentiment analysis

Sentiment Analysis

- Used BERT-based[1] Turkish sentiment model.
- Twitter score is calculated based on the formula below which uses like, quote, reply, retweet and follower counts.

$$z(f) = (f - \mu) / \sigma \quad (1)$$

$$IS(l, RE, RT, f) = log(l) z(f) \sqrt{(RE + RT)} \quad (2)$$

Imputation

- To fill out missing dates without news or tweets we tried multiple imputation are tested.
- Approximately %75 of KAP and %8 of Twitter data was missing.

Model Training

- 80% of data is used for training, remaining %20 is used for testing.
- F1-score is used for accuracy metric
- 12 different machine learning models are tested.[3][4]

EXPERIMENTAL RESULTS

Results using KAP sentiment score

%	2 Bins(Best)	6 Bins(Best)	2 Bins(Mean)	6 Bins(Mean)	Best Model(2 Bins)	Best Model(6 Bins)
AKBNK	77.4	56.1	55.8	34.3	Logistic Regression	Linear SVC
EKGYO	57.9	46.9	45.5	21.4	KNN	KNN
EREGL	60.5	46.9	50.4	27.5	Gradient	Decision Tree
GARAN	67.2	52.8	53.4	30.5	Linear SVC	Naive Bayes
GUBRF	69.2	33.1	53.8	21.6	Linear SVC	Decision Tree
ISCTR	62.9	55.2	50.9	28.8	Logistic Regression	XGBoost
KCHOL	64.2	52.6	53.6	32.8	Linear SVC	Decision Tree
KRDMD	62	41.7	51.5	24	SVC	XGBoost
PETKM	60.4	48.8	51.3	25.3	Linear SVC	XGBoost
TAVHL	69.1	49.5	56	25.8	Naïve Bayes	Random Forest

%	AKBNK	EKGYO	EREGL	GARAN	GUBRF	ISCTR	KCHOL	KRDMD	PETKM	TAVHL
Decay Fill	77.1	55.7	**60.5**	**67.2**	54	57.2	**64.2**	60.2	56.1	**69.1**
Decay Fill All	**77.4**	55.7	58.2	65.8	63.6	56.7	59.5	57.2	56.5	65.8
Iterative Imputation	68.9	53.8	58.1	**67.2**	66.1	**62.9**	60.5	60.4	60.1	64.5
KNN Imputation	72.4	**57.9**	59.9	64.7	64.6	59.9	57.4	59.1	**60.4**	65.3
Mean Imputation	75.8	55.9	57.8	65.7	**69.2**	56.2	62.6	**62**	59.4	68.1
Zero Imputation	76.4	56.2	57	64.9	57.8	54.7	63.1	57.2	55.9	68.4

Results without KAP sentiment score.

%	2 Bins (Best)	6 Bins (Best)	2 Bins (Mean)	6 Bins (Mean)	Best Model (2 Bins)	Best Model (6 Bins)
AKBNK	76.4	56.9	55.7	34.1	Logistic Regression	Linear SVC
EKGYO	55.9	46.9	45.3	21.5	KNN	KNN
EREGL	61	44.7	50	27	XGBoost	Decision Tree
GARAN	65.5	52.6	38.9	29.7	Linear SVC	Naïve Bayes
GUBRF	68.3	33.9	53.2	21.1	Naïve Bayes	Decision Tree
ISCTR	56.7	55.9	48.4	28.4	Decision Tree	Random Forest
KCHOL	64.4	52.3	53.4	33.1	Logistic Regression	Linear SVC
KRDMD	64	42.2	51.5	22.9	Decision Tree	AdaBoost
PETKM	64.3	49.2	50.8	24.5	SVC	XGBoost
TAVHL	68.7	50.5	56.1	25.2	Naïve Bayes	Random Forest

%	AKBNK	EKGYO	EREGL	GARAN	GUBRF	ISCTR	KCHOL	KRDMD	PETKM	TAVHL
Decay Fill	**76.4**	55.9	**61**	**65.5**	66.1	**56.7**	57.1	60.1	57.3	**68.7**
Decay Fill All	76.4	53.8	56	61.4	**68.3**	56.6	56.1	**64**	**64.3**	66.6
Iterative Imputation	69.4	55.2	55.9	64.4	64.4	56.5	59.2	60	59.6	66.8
KNN Imputation	73.7	**55.9**	56.9	62.1	67.3	55.1	57.8	58.6	58.7	67.5
Mean Imputation	76.4	54.2	57.9	64.9	68.1	55.4	61.7	63.6	56	68
Zero Imputation	76.4	54.1	58.2	65	58.2	55.4	**64.4**	60.4	57.7	68.3

REFERENCES

[1] Kilimci, Z. H., & Duvar, R. (2020). An Efficient Word Embedding and Deep Learning Based Model to Forecast the Direction of Stock Exchange Market Using Twitter and Financial News Sites: A Case of Istanbul Stock Exchange (BIST 100). *IEEE Access*, *8*, 188186-188198.
[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.
[3] Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, *42*(22), 9001-9011.
[4] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.

CONCLUSION

- Determined the additional value added by using KAP sentiments.
- Optimized imputation methods to correctly assess the value of sentiments as time passes.
- This is the first project in this field that predicts direction of an **individual** stock by using Twitter and KAP data.
- Our model gives %56 accuracy on average and at best it can predict with %77.4 accuracy for 2 classes.