https://doi.org/10.17576/apjitm-2020-0901-01
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 9 No. 1, June 2020: 1 - 16
e-ISSN: 2289-2192

Received: 28 January 2020
Accepted: 5 May 2020
Published: 15 June 2020

BITCOIN PRICE PREDICTION BASED ON SENTIMENT OF NEWS ARTICLE AND MARKET DATA WITH LSTM MODEL

CHEE KEAN CHIN
NAZLIA OMAR

ABSTRACT

Bitcoin is a digital currency and investment tool that has received worldwide attention recently. However, the fluctuation of Bitcoin price has been a concern to the users and investors. Forecasting the Bitcoin price can serve as a guideline for investor and user to make effective strategy in their investment or usage. With the rapid development of the Internet, online data including news article can facilitate forecasting Bitcoin price. This research aims to study the effect of news article sentiment towards Bitcoin price with study period from September 2017 to August 2019. Accordingly, this study introduces sentiment analysis to understand the relevant information of online news articles and use it as an input feature for Bitcoin price prediction. Two main phases are included in the study, which is sentiment analysis and price prediction. In sentiment analysis, the sentiment is extracted based on a lexicon-based approach to capture the relevant news articles information regarding cryptocurrency markets. In price prediction, the sentiment is used as an input feature and Long Short-Term Memory (LSTM) model is used in price prediction phase. With Bitcoin market data and news articles as samples, the empirical results show that news articles sentiment reduced the overall error in Bitcoin price predictions.

Keywords: Bitcoin, Prediction, LSTM, Lexicon, Sentiment

RAMALAN HARGA BITCOIN BERASASKAN POLARITI SENTIMEN ARTIKEL BERITA DAN DATA PASARAN DENGAN MODEL LSTM

ABSTRAK

Bitcoin adalah wang digital dan alat pelaburan yang telah mendapat perhatian seluruh dunia sejak kebelakangan ini. Namun, harga Bitcoin yang tidak stabil telah menjadi kebimbangan di kalangan pengguna dan pelabur Bitcoin. Ramalan harga Bitcoin dapat membantu pelabur dan pengguna untuk membina strategi yang efektif dalam pelaburan atau penggunaan. Dengan perkembangan pesat Internet, data dalam talian termasuk artikel berita boleh membantu dalam harga ramalan Bitcoin. Kajian ini bertujuan untuk mengkaji kesan sentimen artikel berita kepada harga Bitcoin dengan tempoh kajian dari September 2017 hingga Ogos 2019. Sehubungan dengan itu, kajian ini memperkenalkan analisis sentimen untuk memahami maklumat artikel berita dalam talian dan menggunakannya sebagai fitur input untuk ramalan harga Bitcoin. Terdapat dua fasa utama dalam kajian ini, iaitu analisis sentimen dan ramalan harga. Dalam analisis sentimen, sentimen diekstrak berdasarkan kaedah leksikon untuk memahami maklumat artikel berita berkaitan dengan pasaran kriptowang. Kriptowang adalah sejenis sistem pembayaran digital dan monetari yang mana transaksi dilakukan dengan cara desentralisasi yang merupakan transaksi kewangan rakan-ke-rakan tanpa melalui institusi kewangan. Dengan kata lain, Bitcoin tidak bergantung kepada perantara pihak ketiga untuk memproses pembayaran, ia menggunakan bukti kriptografi dalam komputer untuk memproses dan mengesahkan kesahihan dan menyebarkan antara rangkaian (Nakamoto 2008). Dalam ramalan harga, sentimen digunakan sebagai fitur input dan model Memori Jangka Panjang Pendek (LSTM) digunakan dalam fasa ramalan harga. Dengan data pasaran dan artikel berita

sebagai sampel, keputusan menunjukkan sentimen artikel berita dapat mengurangkan kesilapan dalam ramalan harga Bitcoin.

Katakunci: Bitcoin, Ramalan, LSTM, Leksikon, Sentimen

PENGENALAN

Bitcoin adalah perintis wang kripto dan telah menjadi wang kripto yang paling popular di dunia (Nakamoto 2008). Wang kripto adalah sejenis sistem pembayaran digital dan monetari yang mana transaksi dilakukan dengan cara desentralisasi yang merupakan transaksi kewangan rakan-ke-rakan tanpa melalui institusi kewangan. Dengan kata lain, Bitcoin tidak bergantung kepada perantara pihak ketiga untuk memproses pembayaran, ia menggunakan bukti kriptografi dalam komputer untuk memproses dan mengesahkan kesahihan dan menyebarkan antara rangkaian (Nakamoto 2008). Oleh itu, Bitcoin boleh digunakan sebagai pembayaran yang boleh dibuat dalam talian tanpa memerlukan kawalan dan kos untuk pihak ketiga sebagai perantara yang dipercayai untuk mengesahkan transaksi. Selain itu, Bitcoin juga dianggap sebagai pelaburan alternatif yang baru dalam pasaran kewangan.

Ramalan harga sentiasa menjadi topik penyelidikan yang menarik kerana ramalan yang tepat boleh memberi garis panduan kepada institusi dan individu untuk membuat keputusan pelaburan dan keupayaan untuk merancang dan membangunkan strategi yang berkesan. Berdasarkan kajian lepas bahawa ramalan dapat dilakukan dari dua perspektif: teknik statistik dan pembelajaran mesin (Wang et al. 2012). Model autoregresi bersepadu purata bergerak (ARIMA) adalah model statistik yang amat digunakan dalam peramalan data siri masa (Lee et al. 2007; Merh et al.2010). Dari perspektif pembelajaran mesin, model rangkaian neural buatan (ANN) merupakan model yang sangat popular kerana keupayaan dalam mempelajari corak dari data dan infer penyelesaian daripada data mentah yang tidak diketahui (Nigam 2018). LSTM adalah sejenis rangkaian neural berulang (RNN) yang dapat mempelajari pergantungan pesanan antara item dalam turutan. LSTM direka untuk mempelajari kebergantungan jangka panjang dan mengingati maklumat untuk tempoh yang lama yang sesuai untuk data siri masa di mana analisis siri masa bergantung kepada perubahan nilai masa lalu untuk ramalan nilai masa depan (Orac, 2019).

Untuk meramalkan harga Bitcoin masa depan, faktor asas yang mendorong harga Bitcoin harus ditentukan. Namun, Yermack (2013) mendakwa bahawa Bitcoin tidak mempunyai nilai intrinsik. Dengan kata lain, Bitcoin tidak mempunyai faktor asas dan sukar untuk memberikan nilai asas kepada Bitcoin (Gómez-González et al. 2014), nilai Bitcoin bersifat subjektif dan lebih cenderung kepada pengaruh sentimen pasaran dan sentimen harus dikaitkan dengan pergerakan harga (McAteer 2014). Oleh kerana ciri spekulatif Bitcoin, kebanyakan pelaburan dalam pasaran Bitcoin menilai Bitcoin berasaskan sentimen dan bukannya penilaian nilai ketara aset (Carrera 2018). Berdasarkan kajian lepas, didapati bahawa sebahagian besar penyelidik menggunakan kaedah berasaskan leksikon untuk menganalisis data mentah yang belum dilabelkan dari media sosial seperti Twitter dan menjadikan sebagai peramal harga Bitcoin. Namun, hasil kajian lepas telah menunjukkan keputusan yang tidak selaras dalam menentukan kesan sentimen media sosial terhadap harga Bitcoin.

Selain sentimen daripada media sosial, sentimen berita kewangan dalam talian juga merupakan salah satu faktor dalam talian yang boleh digunakan untuk memperolehi sentimen pasaran Bitcoin. Artikel berita boleh memberi impak yang besar kepada pasaran dan sentimen pelabur, mengakibatkan perubahan dinamik dalam ciri-ciri risiko alam pelaburan (Mitra & Mitra 2012). Dengan itu, berita merupakan salah satu faktor yang penting dalam keputusan pelaburan dan telah digunakan secara meluas dalam ramalan harga atau trend pasaran saham dan menghasilkan hubungan yang penting antara sentimen artikel berita dan pasaran saham

(Chowdhury et al. 2014; Kalyani et al. 2016; Kim et al. 2014; Seng et al. 2017; Shah et al. 2018).

Berdasarkan penyelidikan Bollampelly (2016), 54% pelabur mendapat berita pelaburan dari laman web berita kewangan dan 40% dari laman web media sosial. Ini telah menunjukkan bahawa sentimen artikel berita dari laman web berita kewangan memainkan peranan penting bagi pelabur dalam mendapatkan maklumat tentang pelaburan. Kebanyakan kajian lepas memberi tumpuan kepada sentimen medfia sosial untuk meramalkan harga / trend Bitcoin di mana kesan sentimen artikel berita terhadap harga Bitcoin telah diabaikan. Manakala, kepentingan sentimen artikel berita kewangan dalam pasaran saham telah dibuktikan dalam beberapa kajian lepas (Kalyani et al. 2016; Seng et al.2017; Shah et al. 2018). Ini menunjukkan bahawa selain dari pada sentimen media sosial, sentimen dari media lain seperti artikel berita yang menunjukkan pengaruhan signifikan terhadap pasaran saham boleh dijadikan salah satu faktor yang penting terhadap pasaran Bitcoin seperti dalam pasaran saham. Oleh itu, kajian ini akan mengaji kesan artikel berita terhadap harga Bitcoin.

KAJIAN LEPAS

Kebelakangan ini, Bitcoin sebagai pelaburan baru yang bersifat spekulasi telah menarik perhatian penyelidik-penyelidik untuk memahami faktor-faktor yang mempengaruhi harga trend Bitcoin. Oleh itu, penyelidik telah berusaha dalam pembangunan model-model yang dapat meramalkan trend masa depan pasaran Bitcoin. Kebanyakan teknik yang sedia ada menggunakan data pasaran, open, high, low, close, dan volume (OHLCV) dan sentimen daripada media sosial sebagai peramal Bitcoin. Namun, kajian lepas menghasilkan keputusan yang tidak konsisten. Beberapa penyelidik menunjukkan bahawa terdapat hubungan antara sentimen media sosial dan turun naik harga Bitcoin, begitu juga sebaliknya. Selain daripada sentimen media sosial, beberapa kajian lepas mendapati bahawa sentimen artikel berita mempunyai pengaruh yang kuat terhadap trend pasaran saham dan boleh digunakan sebagai penunjuk trend harga. Berikut adalah perbincangan kajian lepas yang mengenai ramalan harga Bitcoin dengan menggunakan analisis sentimen dan kesan artikel berita kepada pasaran kewangan.

Stenqvist et al. (2017) telah membangunkan model untuk meramalkan perkembangan harga Bitcoin befrdasarkan Twitter, 2,271,815 twit dan seterusnya menganalisis sentimen dengan menggunakan kaedah leksikon VADER. Markah sentimen kompaun dengan ambang 0.5 dan dibandingkan dengan harga Bitcoin. Pengesanan kajian menyimpulkan bahawa sentimen boleh meramalkan arah pembangunan harga Bitcoin dengan ketepatan sehingga 79%. Galeshchuk et al. (2018) menggunakan analisis sentimen untuk meramalkan pergerakan harga Bitcoin dengan menggunakan data dari Twitter dan menjalankan analisis sentimen berasaskan leksikon dengan menggunakan leksikon PATTERN, python leksikon terbina dalam. Kajian beliau menunjukkan bahawa terdapat pengaruh yang signifikan dari sentimen Twitter kepada turun naik harga Bitcoin. Sementara itu, terdapat juga satu kajian yang menunjukkan kesan skor sentimen media sosial dari Twitter kepada harga Bitcoin dengan data berlabel secara manual (Pant et al. 2018). Keputusan kajian mencadangkan dengan mengambil kira harga lepas dan skor sentimen twit positif dan negatif boleh menghasilkan hasil yang lebih baik.

Walau bagaimanapun, terdapat beberapa kajian yang sedia ada menunjukkan percanggahan hasil yang menolak kesan sentimen media sosial kepada pasaran Bitcoin. Abraham et al. (2018) telah mengkaji ramalan harga Bitcoin dengan analisis sentimen berdasarkan sentimen dan volum twit, 30,420,063 twit telah dikumpul dalam tempoh masa 60 hari. Beliau mencadangkan bahawa sentimen Twitter tidak konsisten dengan perubahan harga Bitcoin dan tidak mempunyai hubungan yang jelas antara sentimen twit dan harga. Salač (2019) juga telah menjalankan analisis sentimen berdasarkan data Twitter dan Reddit

sepanjang tempoh 92 hari, sebanyak 999,879 Twits dan 124,681,323 komen Reddit telah dihasilkan dan dinilai dengan sentimen leksikon VADER, namun kajian ini menyimpulkan bahawa hubungan penyebab tidak dapat dibuktikan. Satu lagi kajian tentang hubungan antara harga Bitcoin dan sentimen Twitter yang dijalankan oleh Carrera (2018) menyimpulkan bahawa sukar untuk menentukan secara konklusif jika hubungan yang signifikan wujud antara sentimen awam dan harga Bitcoin.

Selain daripada sentimen media sosial, beberapa kajian lepas juga menggunakan sentimen artikel berita kewangan untuk ramalan harga dalam pasaran kewangan. Chowdhury et al. (2014) menjalankan kajian analisis sentimen berita untuk meramalkan trend harga saham dengan menggunakan analisis sentimen mengenai tajuk berita dan siaran akhbar yang berkaitan berdasarkan pendekatan berasaskan leksikon. Kajian beliau menyimpulkan terdapat korelasi yang kuat antara sentimen berita dan harga saham. Ini menunjukkan bahawa sentimen dapat mencerminkan pergerakan harga saham. Kajian serupa telah dilakukan oleh Kalyani et al. (2016) di mana analisis sentiment dijalankan berdasarkan artikel berita dengan menggunakan kaedah leksikon untuk mengkaji hubungan antara artikel berita dan trend saham dan seterusnya meramalkan harga pasaran saham. Kajian ini menunjukkan bahawa aliran saham boleh diramalkan dengan menggunakan artikel berita dan harga saham terdahulu. Di samping itu, terdapat beberapa kajian yang sedia ada sejajar dengan hasil yang mencadangkan hubungan yang signifikan secara statistik antara berita kewangan dan volatiliti pasaran saham (Kim et al. 2014; Seng et al. 2017).

Dalam kajian (Mcnally 2016.; Saxena et al. 2018), mereka menggunakan model pembelajaran mendalam, LSTM dan model statistik tradisional, ARIMA untuk meramalkan harga Bitcoin untuk membandingkan ramalan ketepatan kedua-dua model. Keputusan menunjukkan bahawa model LSTM telah mengalahkan model ARIMA dalam kajian mereka. Dalam kajian Mcnally (2016), model LSTM mencapai RMSE 8% di mana lebih kurang daripada RMSE 53% yang dihasilkan oleh model ARIMA. Kajian yang telah dijalankan oleh Saxena et al. (2018), RMSE model ARIMA adalah 700.69 di mana lebih tinggi berbanding dengan RMSE model LSTM iaitu 456.78. Ini menunjukkan prestasi model LSTM adalah lebih baik dalam ramalan harga Bitcoin berbanding dengan model ARIMA.

Dengan menggunakan ringkasan kajian sedia ada, rangka kerja boleh dirangka. Kesimpulannya, kesan sentimen media sosial terhadap harga Bitcoin dalam kajian lepas mempunyai hasilan yang berbeza, manakala tiada kajian mengambil kira sentimen artikel berita sebagai peramal harga Bitcoin seperti yang digunakan dalam pasaran saham di mana telah menunjuk hubungan yang statistik signifikan dalam pasaran saham.

METODOLOGI

Bahagian ini akan membincangkan proses-proses pembangunan model ramalan harga Bitcoin. Rajah 1 dan Rajah 2 mempamerkan rangka kerja keseluruhan kajian dengan bahagian seperti pengumpulan data, analisis sentimen, normalisasi data dan sebagainya.

PENGUMPULAN DATA

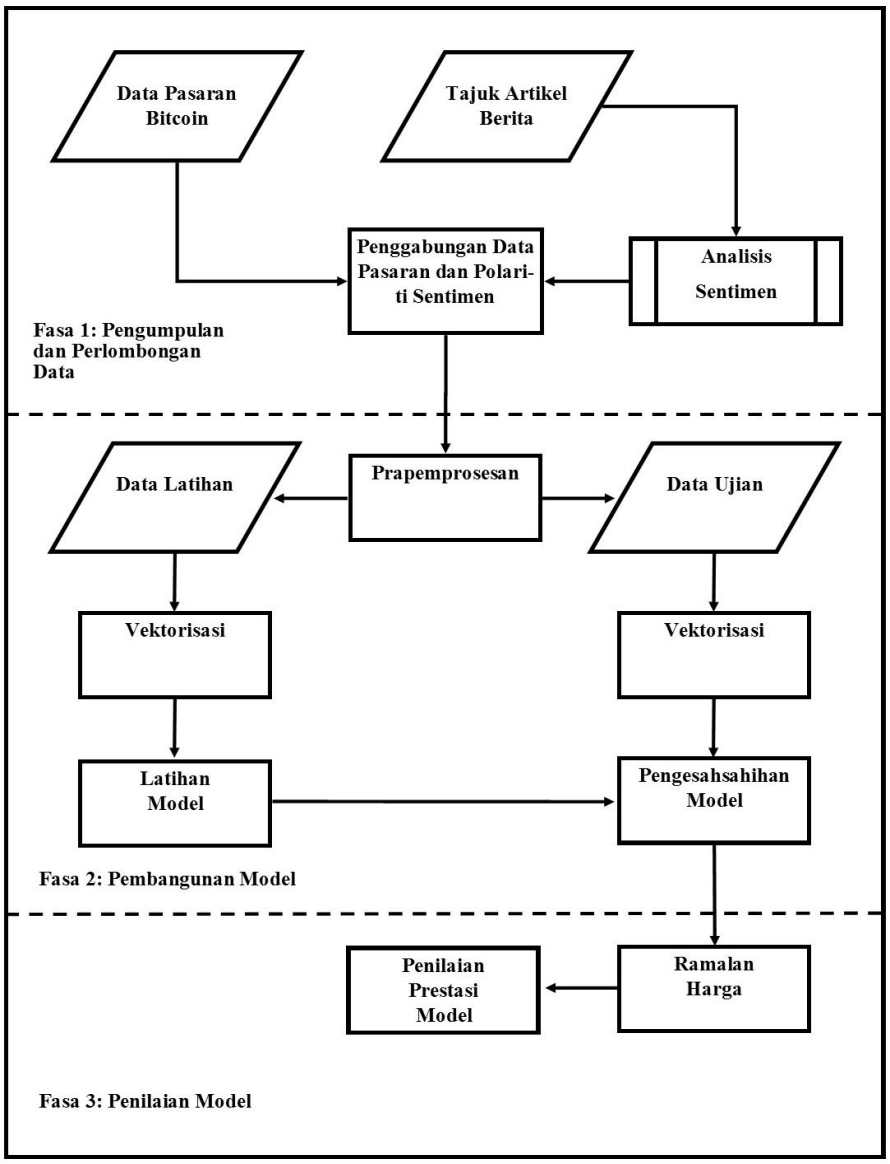
Sumber input tajuk artikel berita adalah dari platform pasaran kewangan (investing.com) yang menyediakan data masa nyata, carta, alat kewangan, berita gempar dan analisis pasaran kewangan. Rangka masa bagi pengumpulan data berita adalah bermula dengan berita yang berkaitan dengan wang kripto yang terawal dan berakhir pada September 2019. Dengan itu, sejumlah 15,295 artikel telah dikumpul dari laman web ini. Selain itu, data pasaran Bitcoin juga dikumpulkan dengan tempoh masa dan sumber yang sama. Data pasaran mengadungi 6 fitur iaitu, Harga, Open, High, Low, dan Volum. Harga dan Open adalah harga terakhir dan

harga bermula harian. High dan Low mewakili pencapaian harga tertinggi dan harga terendah harian. Volum merupakan jumlah perdagangan harian dalam pasaran.

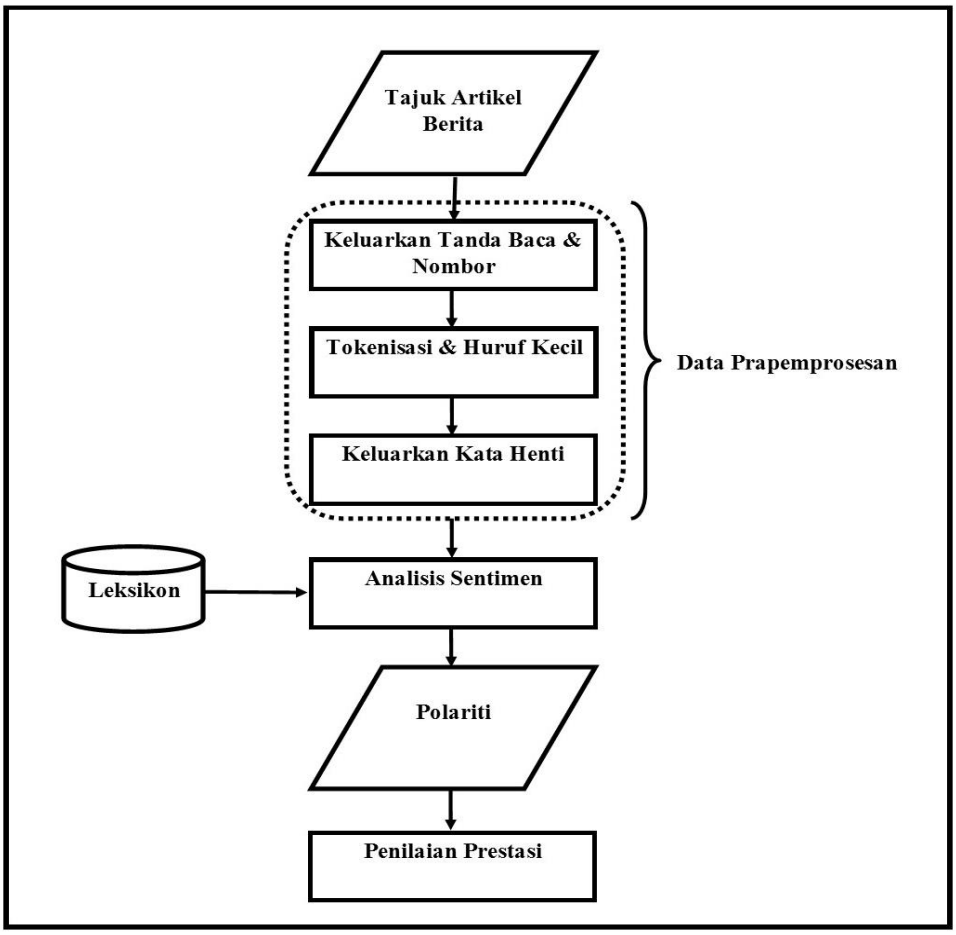
Dalam pembelajaran mesin, data perlu diasingkan kepada data latihan dan ujian bagi fasa latihan dan ujian yang berasing. Sebelum model dapat menjalankan ramalan, model akan berlajar daripada data latihan. Dalam proses ini, data akan dipisahkan kepada dua kumpulan iaitu, data latihan dan data ujian. Data yang dikumpulkan daripada 10 September 2017 hingga 15 Ogos 2019 mempunyai 700 bilangan dalam bentuk harian. Tempoh dipilih dalam kajian ini adalah kerana ketersediaan data dalam talian. Dengan 700 data ini, tiga dataset yang mempunyai saiz latihan dan ujian yang berbeza telah dibentukkan untuk peramalan harga Bitcoin. Tujuan utama tiga dataset berbeza diadakan dalam saiz latihan dan ujian adalah untuk menunjukkan walaupun dengan rangka masa yang berbeza ralat peramalan juga dapat dikurangkan dengan melibatkan sentimen artikel berita dan bukannya kesekenaan. Struktur dataset ditunjuk dalam Rajah 3.

NORMALISASI TEKS DATA

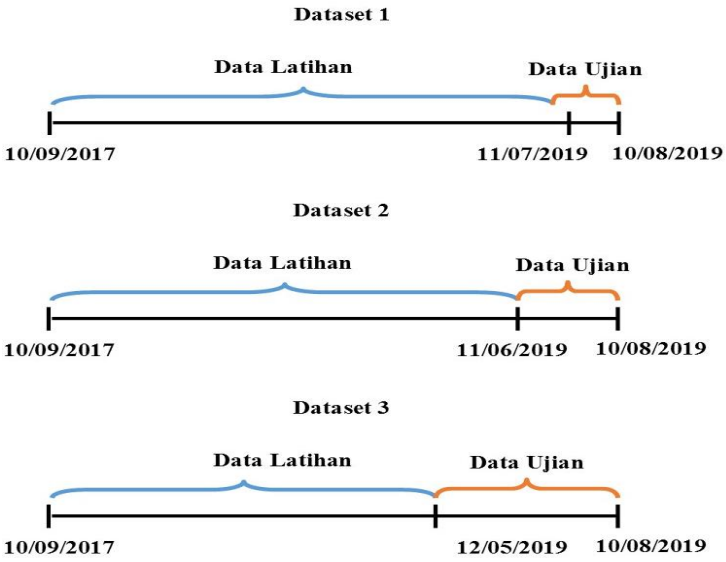
Dalam proses ini, pembersihan data dilakukan untuk memastikan bahawa data teks dinormalkan dan pengurangan hingar akan dilakukan. Pertama, artikel duplikasi telah dibuang kerana ia tidak memberi informasi yang berguna dan akan mempengaruhi keputusan ramalan harga kemudian. Kedua, langkah teks data normalisasi telah dilakukan untuk mengurangkan hingar dalam data seperti, keluarkan tanda baca, keluarkan nombor, mengasingkan data dalam bentuk token, ubah perkataan ke dalam huruf kecil dan keluarkan kata henti. Rajah 4 merupakan contoh tajuk artikel berita selepas normalisasi teks.



RAJAH 1. Rangka Kerja Pembangunan Model



RAJAH 2. Proses Analisis Sentimen



RAJAH 3. Data Latihan dan Ujian

Tajuk Artikel Berita	Selepas Prapemprosesan
$3B Ponzi Scheme Is Now Allegedly Dumping Bitcoin by the Hundreds	['ponzi', 'scheme', 'allegedly', 'dumping', 'bitcoin', 'hundreds']
Bitcoin Price Falls Suggest Bubble is Bursting, Economists Warn	['bitcoin', 'price', 'falls', 'suggest', 'bubble', 'bursting', 'economists', 'warn']
Bitcoin (BTC) Too Volatile for Its Own Good; Volatility May Hurt ETF Efforts	['bitcoin', 'btc', 'volatile', 'good', 'volatility', 'may', 'hurt', 'etf', 'efforts']
Use of Bitcoin for Payments Declines Despite Its Bigger Stability - Report	['use', 'bitcoin', 'payments', 'declines', 'despite', 'bigger', 'stability', 'report']
Bitcoin Climbs Higher Despite Increasing Concerns of Imminent Correction	['bitcoin', 'climbs', 'higher', 'despite', 'increasing', 'concerns', 'imminent', 'correction']
Bitcoin Remains Bearish as Cryptos Fall	['bitcoin', 'remains', 'bearish', 'cryptos', 'fall']
Bitcoin Falls 10% In Bearish Trade	['bitcoin', 'falls', 'bearish', 'trade']
Bitcoin Gains Despite BOJ Comes Up with Negative Q&A	['bitcoin', 'gains', 'despite', 'boj', 'comes', 'negative']
Bitcoin Gains Despite Worst December in 7 Years	['bitcoin', 'gains', 'despite', 'worst', 'december', 'years']
Bitcoin Falls Under $8,000 Again as US Stock Market Sees Discrete Gains	['bitcoin', 'falls', 'us', 'stock', 'market', 'sees', 'discrete', 'gains']
Bitcoin Falls Near $9,000 as US Stock Market Sees Gains	['bitcoin', 'falls', 'near', 'us', 'stock', 'market', 'sees', 'gains']

RAJAH 4. Contoh Tajuk Artikel Berita Selepas Prapemprosesan.

ANALISIS SENTIMEN

Kaedah berasaskan leksikon menentukan orientasi sesuatu dokumen dengan menilai perkataan yang ditulis terhadap sentimen dalam leksikon. Kedua-dua Loughran dan Mcdonald (2011) dan Siering (2012) menunjukkan bahawa prestasi meningkat secara drastik apabila menggunakan leksikon yang spesifik. Leksikon biasa terbukti lebih kerap salah mengenal pasti konotasi atau konteks sesuatu perkataan (Soroka et al. 2015). Oleh itu, leksikon yang fokus pada bidang kewangan yang dicipta oleh Loughran dan Mcdonald (2011) dan Chen et al. (2018) akan digunakan dalam kajian ini. Prestasi leksikon diukur dari segi ramalan yang betul akan dibuat.

Leksikon Loughran dan McDonald (2011) adalah salah satu leksikon popular dalam bidang kewangan. Leksikon ini mempunyai sejumlah 2,709 perkataan, dengan 354 perkataan adalah positif dan 2,355 perkataan adalah negatif. Berdasarkan leksikon, sentimen boleh dikira menggunakan rumus di dalam (1) berikut:

$$s_t = \frac{p_p(T_t) - p_n(T_t)}{p_p(T_t) + p_n(T_t)}, \quad (1)$$

Dengan $\boldsymbol{T_t}$ adalah tajuk artikel berita yang sedia ada pada masa t, $\boldsymbol{p_p(T_t)}$ adalah jumlah bilangan perkataan positif dalam $\boldsymbol{T_t}$, $\boldsymbol{p_n(T_t)}$ adalah jumlah bilangan perkataan negatif, dan $\boldsymbol{s_t}$ adalah sentimen yang sepadan. Jelas sekali, $\boldsymbol{s_t}$ [-1,1] adalah perbezaan antara bilangan positif dan negatif perkataan yang dibahagikan dengan jumlah perkataan positif dan negatif dalam tajuk artikel berita yang ada pada $\boldsymbol{T_t}$.

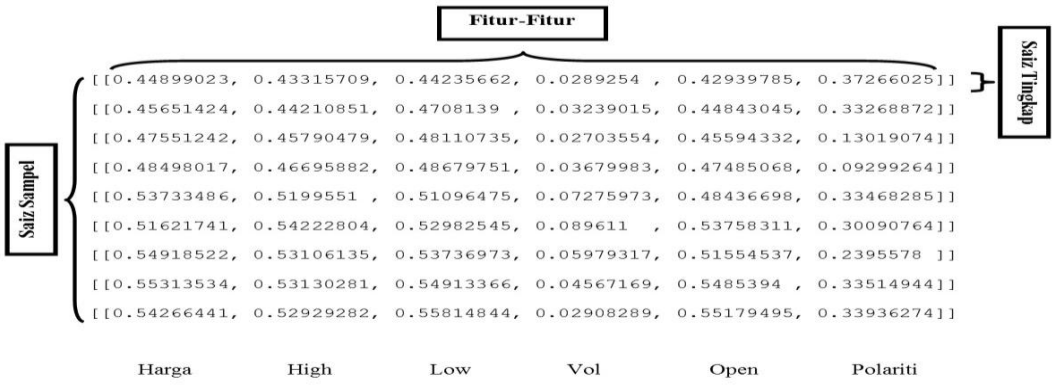
Leksikon NTUSD dibina oleh Chen et al. (2018) yang mengumpulkan sejumlah 8,331 perkataan dan nilai sentimen bagi setiap perkataan dengan nilai -3.812 hingga 1.222. Apabila perkataan muncul dalam leksikon, nilai sentimen yang diberi kepada perkataan tertentu akan ditambah dan mengembalikan jumlah skor untuk satu dokumen. Oleh itu, setiap dokumen akan mempunyai skor sentimen sendiri dan akhirnya semua skor sentimen pada hari yang sama akan digabungkan menjadi satu skor. Polariti yang dijana oleh leksikon yang mempunyai prestasi yang terbaik dalam pengelasan akan digunakan sebagai fitur input bagi peramalan harga Bitcoin.

NORMALISASI DATA

Pembelajaran mesin mempelajari cara pemetaan input kepada output daripada dataset latihan. Oleh kerana penggunaan pemberat kecil dalam model dan penggunaan kesilapan antara ramalan dan nilai cerapan, skala input dan output yang digunakan untuk melatih model menjadi faktor yang penting. Input tanpa normalisasi akan menjadikan proses pembelajaran model belajar secara perlahan atau tidak stabil, bagi output yang tanpa normalisasi dalam masalah regresi boleh mengakibatkan isu kecerunan sehingga proses pembelajaran gagal. Penskalaan fitur (juga dikenali sebagai normalisasi data) ialah kaedah yang digunakan untuk menyeragamkan pelbagai ciri data. Matlamat normalisasi adalah untuk menukar nilai angka dalam dataset ke skala yang sama, tanpa mengubah perbezaan dalam julat nilai-nilai. Dalam penskalaan (juga dikenali sebagai skala min-max), data akan diubah supaya ciri-ciri tersebut berada dalam julat tertentu misalnya, 0 hingga 1. Dalam proses ini, data skor sentimen dan data pasaran Bitcoin telah digabungkan menjadi satu dataset sebelum proses normalisasi dijalankan.

VEKTORISASI DATA

Reka bentuk input dalam lapisan input LSTM perlu ditentukan dalam vektor tiga dimensi spesifik: Sampel, Saiz tingkap, dan Bilangan fitur. Sebagai contoh, bentuk input data Bitcoin yang digunakan dalam model latihan adalah: (668, 1, 6), maksudnya dalam vektor ini mempunyai 668 sampel, saiz tingkap (“timestep”) 1, dan 6 fitur. Dalam kajian ini saiz tingkap 1 digunakan kerana ramalan berdasarkan maklumat satu hari sebelumnya. Akhirnya, output vektor yang dijana dalam proses ini merupakan vektor data input yang akan digunakan dalam lapisan input model LSTM. Rajah 5 menunjukkan contoh output vektor.



RAJAH 5. Contoh Vektor Data.

PERAMALAN HARGA

Model seperti ARIMA bergantung kepada andaian data linear. Oleh kerana harga Bitcoin merupakan data siri masa yang tak linear, model ini tidak dapat memberikan hasil yang berguna (Okasha et al. 2016). Berdasarkan, harga Bitcoin yang tak linear dan kesesuaian pembelajaran mendalam dalam ramalan data siri masa, maka kajian ini menggunakan teknik pembelajaran mendalam, khususnya model LSTM, untuk meramalkan harga Bitcoin. LSTM mempunyai keupayaan untuk memproses data dan menghantarkan maklumat sel sebelumnya ke seluruh rangkaian LSTM.

Untuk membina model LSTM yang dapat meramal harga Bitcoin, satu set parameter telah dipilih untuk mengoptimumkan proses latihan dan meminimumkan kesilapan. Dalam kebanyakan kes, kajian lepas mencadangkan penggunaan kaedah cuba jaya dalam menentukan parameter rangkaian. Keturunan Gradien Stochastic banyak digunakan dalam Rangkaian Neural, ia mempunyai masalah menumpu kepada minimum. Beberapa pengoptimuman yang popular adalah variasi algoritma pembelajaran adaptif, seperti Adam, Adagrad, dan RMSProp.

Kingma dan Lei Ba (2014) menunjukkan bahawa pembetulan bias membantu Adam mengalahkan RMSprop dan Adagrad dalam pengerakan ke arah akhir pengoptimuman kerana kecerunan menjadi lebih sparser. Oleh itu, Adam akan digunakan sebagai pengoptimum dalam kajian ini.

Fungsi Pengaktifan yang paling popular adalah sigmoid, tanh, dan ReLu. Tanh dan Sigmoid mempunyai masalah kecerunan lenyap, di mana rangkaian neural berhenti belajar atau dengan secara yang perlahan. Namun, ReLu tidak mempuyai isu kecerunan lenyap dan ia memerlukan kuasa pemprosesan yang lebih ringan berbanding tanh dan sigmoid kerana ia melibatkan operasi matematik yang lebih mudah (Ruder 2016). Di antara kajian lepas, Wang et al. (2005) dan Rene et al (2013) menggunakan kaedah cuba jaya untuk mengkonfigurasi parameter rangkaian untuk simulasi mereka. Oleh itu, dalam kajian ini parameter tertentu juga dipilih dan ditetapkan dengan percubaan. Parameter yang digunakan dalam model ramalan telah dimasuk dalam Jadual 1.

JADUAL 1. Parameter LSTM Model

PARAMETER	Nilai
Pengoptimum	Adam
Fungsi Pengaktifan	ReLu
Drop rate	0.2
Epok	200
Saiz Batch	100
Lapisan tersembunyi	2

PENILAIAN PRESTASI

Siri masa biasanya fokus dalam ramalan nilai sebenar, ataupun masalah regresi. Oleh itu, pengukuran prestasi dalam kajian ini akan memberi tumpuan kepada kaedah-kaedah yang menilai ramalan dalam nilai sebenar. Kejituan tidak boleh digunakan sebagai penilai dalam kajian ini kerana ia adalah penilaian bagi masalah klasifikasi dan bukannya masalah regresi. Pengukuran yang biasa digunakan seperti min ralat mutlak (MAE) dan ralat punca min kuasa dua (RMSE). MAE dikira dengan purata perbezaan mutlak antara nilai sasaran dan ramalan. Salah satu perbezaan antara MAE dan RMSE adalah RMSE mempunyai keupayaan dalam pengendalian ralat besar. Dengan kata lain, RMSE lebih sensitif kepada nilai pesisih (Drakos, 2018). Oleh itu, RMSE akan digunakan untuk mengukur prestasi model ramalan dalam kajian ini. RMSE di dalam (2) adalah punca kuasa dua terhadap perbezaan antara nilai ramalan (P_i) dan nilai cerapan (O_i) dibahagikan dengan bilangan data (n):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}, \qquad (2)$$

ANALISIS DATA

Analisis sentimen dalam kajian ini dijalankan dengan menggunakan kaedah leksikon. Setiap tajuk artikel berita akan diberikan skor sentimen mengikut sentimen perkataan yang disenaraikan dalam leksikon.

Untuk mengesahkan prestasi dua-dua leksikon dalam analisis sentimen, 120 artikel berita dipilih secara rawak dan dilabel secara manual sama ada positif, negatif atau neutral. Dalam proses ini, bahawa nilai polariti yang lebih daripada 0 adalah positif, nilai kurang daripada 0 adalah negatif dan nilai sama dengan 0 adalah neutral. Kejituan kaedah dikira dengan menjumlahkan angka pada matriks kekalutan (Jadual 2 & 3) dan membahagikan

dengan saiz sampel, iaitu 120. Pada akhirnya, leksikon NTUSD mencapai kejituan 65.83% di mana lebih daripada leksikon Loughran dan McDonald (2011) yang mencapai kejituan 35%. Ini telah menunjukkan bahawa prestasi leksikon NTUSD adalah lebih baik berbanding dengan leksikon Loughran dan McDonald (2011). Oleh itu, leksikon NTUSD akan digunakan dalam kajian ini dan sentimen yang dijana akan digunakan sebagai fitur input bagi peramalan harga Bitcoin. Kejituan analisis kaedah ini mungkin boleh mencapai ketepatan yang lebih baik dengan memperluaskan saiz sampel berlabel. Namun, tugas ini memerlukan masa yang panjang dan boleh dikenakan bias manusia.

Jadual 4 menunjukkan keputusan ramalan harga Bitcoin tiga dataset dengan model yang sama. Ia menunjukkan bahawa keputusan yang diramalkan menggunakan model dengan fitur sentimen mempunyai hasil yang lebih baik berbanding dengan model tanpa sentimen. Dalam dataset 1, RMSE yang dijana dengan data pasaran sahaja adalah 4.91% di mana lebih daripada hasil yang dijana dengan termasuk data sentimen iaitu 4.71%. Situasi yang sama dapat dilihat dalam keputusan dataset 2 dan 3, di mana keputusan yang dihasil dengan data pasaran dan data sentimen mempunyai ralat yang lebih kurang, sebanyak 0.20% dan 0.06% kurang.

Plot garis diagnostik berdasarkan sejarah fungsi kerugian tiga model LSTM dalam latihan dan ujian ditunjukkan dalam rajah 6, 7 dan 8. Hasilnya menunjukkan bahawa tiga-tiga model yang merangkumi data sentimen dijalankan dengan prestasi yang ideal, di mana masalah "overfitting" dan "underfitting" tidak wujud. Berdasarkan fungsi kerugian model, ralat ramalan bergerak stabil selepas 12 epok.

JADUAL 2. Matriks Kekalutan (Loughran dan McDonald)

		Kelas Ramalan		
		Negatif	Neutral	Positif
Kelas Benar	Negatif	10	21	1
	Neutral	2	24	1
	Positif	2	51	8

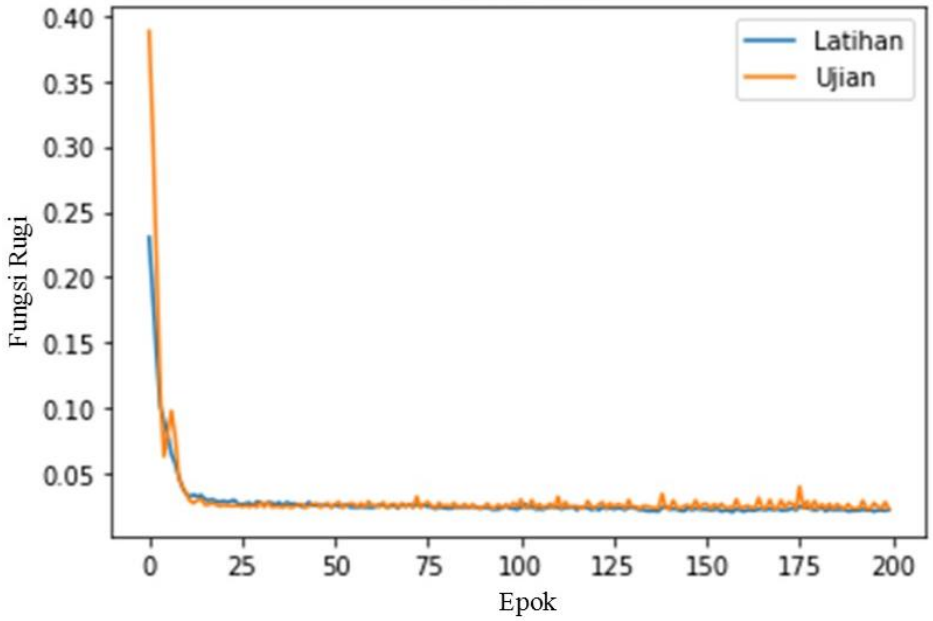
JADUAL 3. Matriks Kekalutan (NTUSD)

		Kelas Ramalan		
		Negatif	Neutral	Positif
Kelas Benar	Negatif	18	0	14
	Neutral	2	4	21
	Positif	4	0	57

JADUAL 4. Keputusan Ramalan Harga Bitcoin

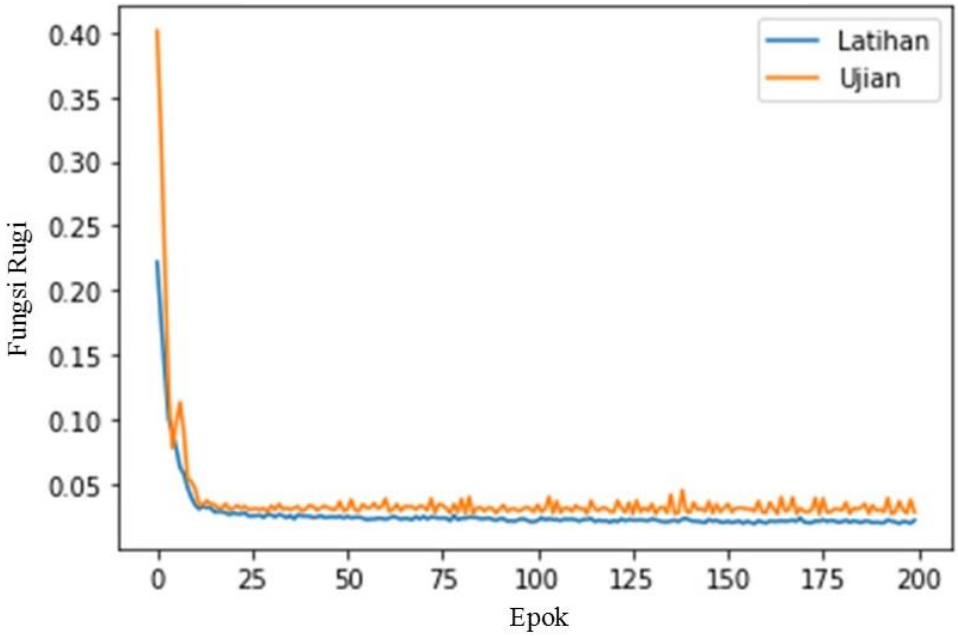
Ramalan	Data Pasaran		Data Pasaran + Sentimen		Ralat
	RMSE	RMSE %	RMSE	RMSE %	
1 (30 hari)	498.76	4.91	484.30	4.71	↓0.20
2 (60 hari)	589.60	5.35	586.56	5.04	↓0.21
3 (90 hari)	517.25	5.04	511.22	4.98	↓0.06

Fungsi Kerugian Dataset 1



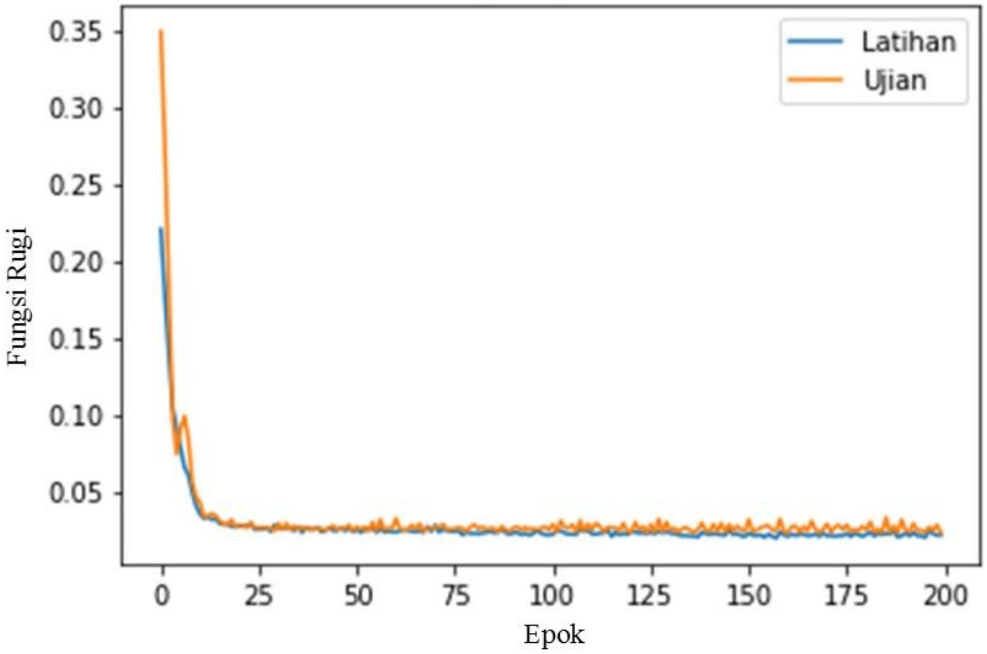
RAJAH 6. Fungsi Kerugian Dataset 1.

Fungsi Kerugian Dataset 2



RAJAH 7. Fungsi Kerugian Dataset 2.

Fungsi Kerugian Dataset 3



RAJAH 8. Fungsi Kerugian Dataset 3.

PERBINCANGAN KEPUTUSAN

Berdasarkan kajian yang dijalankan, keputusan kajian menunjukan sentimen tajuk artikel berita merupakan salah satu fitur yang boleh membantu untuk mengurangkan ralat dalam ramalan harga Bitcoin bagi hari seterusnya walaupun hasilnya tidak menunjuk perbezaan yang besar. Dalam kajian Hileman & Rauchs (2017), ia menunjukkan pengguna dan pelabur Bitcoin berada di seluruh dunia dan bukannya terhad kepada sesuatu kawasan. Oleh kerana, bukan semua pengguna atau pelabur Bitcoin membaca media Inggeris, sentimen tajuk artikel berita dalam bahasa Inggeris telah menjadi salah satu faktor yang menyebabkan kurang pengaruhan dalam ramalan harga Bitcoin.

KESIMPULAN

Hasilan model yang dijanakan dengan data sentimen tajuk artikel berita telah menunjukkan pengurangan ralat berbanding dengan data tanpa sentimen tajuk artikel berita. Walaupun perbezaan hasilnya tidak besar, ia masih mencadangkan nilai yang diramalkan dengan data pasaran dan sentimen tajuk artikel berita lebih tepat dan hampir dengan harga sebenar.

Berdasarkan kepada fungsi kerugian model latihan dan ujian, model yang dibina dengan tiga dataset yang mempunyai saiz latihan dan ujian yang berbeza telah dijalankan dalam keadaan yang ideal, di mana ia tidak mempunyai masalah “overfitting” atau “underfitting” berlaku dalam model tersebut. Sebagai kesimpulan, dengan analisis yang dilakukan, kajian ini boleh menunjukkan bahawa sentimen artikel berita boleh menjadi sebagai faktor dalam ramalan harga Bitcoin.

Dalam kajian ini, leksikon domain kewangan telah digunakan di mana skor sentimen perkataan relevan telah dinyatakan dalam leksikon. Leksikon untuk mendapatkan sentimen berkaitan artikel berita dalam talian, merupakan faktor yang paling penting dalam menentukan polariti sebelum dijadikan sebagai fitur input untuk peramalan harga. Oleh itu leksikon yang

sesuai, terutamanya spesifik untuk pasaran Bitcoin, boleh memberi polariti yang lebih tepat dan keputusan ramalan harga yang lebih sejajar dengan harga benar dapat dijanakan dengan menggunakan polariti tersebut sebagai input.

Berdasarkan kajian Hileman & Rauchs (2017), bilangan orang yang menggunakan kriptowang hari ini telah menyaksikan pertumbuhan yang ketara di seluruh dunia. Oleh itu, kajian masa depan boleh memberi tumpuan kepada kausaliti dengan mengenal pasti sama ada liputan media mempunyai kesan terhadap harga Bitcoin. Ini boleh dilakukan dengan analisis sentimen artikel berita dalam pelbagai bahasa kerana bukan semua pelabur Bitcoin membaca media Inggeris.

RUJUKAN

Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan 2018. "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: 1(3), Article 1. https://scholar.smu.edu/datasciencereview/vol1/iss3/1 [16 October 2019].

Bollampelly, N. R. 2016. Understanding Role of Social Media in Investor Reactions. MBA Dissertation. Dublin Business School.

Carrera, B. P. 2018. Effect of sentiment on bitcoin price formation (pp. 1–49). North Carolina: Duke University Durham.

Chen, C. C., Huang, H. H., & Chen, H. H. 2018. *NTUSD-Fin:* A Market Sentiment Dictionary for Financial Social Media Data Applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*. pp. 37-43

Chowdhury, S. G., Routh, S., & Chakrabarti, S. 2014. News analytics and sentiment analysis to predict stock price trends. *International Journal of Computer Science and Information Technologies* 5(3):3595-3604.

Drakos, G. 2018. How to select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics. https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0 [13 October 2019].

Galeshchuk, S., Vasylchyshyn, O. & Krysovatyy, A. 2018. Bitcoin response to twitter sentiments. *CEUR Workshop Proceedings* 210.: pp. 160–168.

Gómez-González, J. E. & Parra-Polania, J. A. 2014. Bitcoin: something seems to be 'fundamentally' wrong. *Borradores de Economía*; No. 819.

Kalyani, J., Bharathi, N., & Jyothi, R. 2016. Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology. 8. 67-76. 10.5121/ijcsit.2016.8306.

Kim, Y., Jeong, S. R. & Ghani, I. 2014. Text Opinion Mining to Analyze News for Stock Market Prediction. *Int. J. Advance. Soft Comput. Appl* 6(1):1–13

Kingma, D. P. & Lei Ba, J. 2014. Adam: A Method For Stochastic Optimization. Cornell University. https://arxiv.org/pdf/1412.6980.pdf [18 October 2019]

Lee, K., Yoo, S. & Jin, J. J. 2007. Neural Network Model vs. SARIMA Model In Forecasting Korean Stock Price Index (KOSPI). *International Association of Computer Investigative Specialists* 8(2):372-378.

Loughran, T. & Mcdonald, B. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66(1):35-65

McAteer, C. 2014. Twitter Sentiment Analysis to Predict Bitcoin Exchange Rate. Msc Dissertation, University of Dublin.

Mcnally, S. 2016. Predicting the price of Bitcoin using Machine Learning. Masters thesis. Dublin, National College of Ireland.

Merh, N. & Saxena Vinod, P. 2010. A Comparison between Hybrid Approaches of ANN and ARIMA for Indian Stock Trend Forecasting. *Raj Pardasani K. Business Intelligence Journal-July* 3:23-44.

Mitra, L. & Mitra, G. 2012. Applications of News Analytics in Finance: A review. *The Handbook of News Analytics in Finance,* 1–39. John Wiley and Sons.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System [6 August 2019].

Nigam, V. 2018. Understanding Neural Networks. From neuron to RNN, CNN, and Deep Learning. https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90 [5 September 2019].

Orac, R. 2019. LSTM for time series prediction - Towards Data Science. https://towardsdatascience.com/lstm-for-time-series-prediction-de8aeb26f2ca [12 October 2019].

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K. & Lama, B. K. 2018. Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. *IEEE 3rd International Conference on Computing, Communication and Security, ICCCS 2018*. pp. 128–132.

Rene, E. R., López, M. E., Kim, J. H. & Park, H. S. 2013. Back propagation neural network model for predicting the performance of immobilized cell biofilters handling gas-phase hydrogen sulphide and ammonia. *BioMed Research International* 2013. Article ID 463401. pp. 1-10.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 [2 November 2019].

Salač, A. 2019. Forecasting of the cryptocurrency market through social media sentiment analysis. Bsc Dissertation, University of Twente.

Saxena, A. & Sukumar, T. R. 2018. Predicting bitcoin price using lstm And Compare its predictability with arima model. *International Journal of Pure and Applied Mathematics* 119(17): 2591-2600

Seng, J. L. & Yang, H. F. 2017. The association between stock price volatility and financial news – a sentiment analysis approach. *Kybernetes* 46(8): 1341–1365.

Shah, D, Isah, H. & Zulkernine, F. 2018. Predicting the Effects of News Sentiments on the Stock Market. *IEEE International Conference on Big Data* . pp. 4705-4708.

Wang, S., Yu, L. & Lai, K. K. 2005. Crude Oil Price Forecasting With Tei@i Methodology. *Journal of Systems Sciences and Complexity* 18(2):145-166

Siering, M. 2012. Using text mining and sentiment analysis to support intraday investment decisions. *Proceedings of the Annual Hawaii International Conference on System Sciences*. pp. 1050–1059.

Soroka, S., Young, L. & Balmas, M. 2015. Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science* 659(1): 108–121.

Stenqvist, E. & Lönnö, J. 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis. Master Thesis. KTH Royal Institute of Technology.

Wang, J. J., Wang, J. Z., Zhang, Z. G. & Guo, S. P. 2012. Stock index forecasting based on a hybrid model. *Omega* 40(6): 758–766.

Okasha, M., Yaseen, A. 2016. Comparison between ARIMA Models and Artificial Neural Networks in Forecasting Al-Quds indices of Palestine Stock Exchange Market. Conference: The 25th Annual International Conference on Statistics and Modeling in Human and Social Sciences. Cairo University. Vol. 25.

Yermack, D. 2013. Is Bitcoin a Real Currency? An economic appraisal. Working paper in National Bureau of Economic Research. NBER Working Paper No. 19747.

Chee Kean Chin
Nazlia Omar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia
p97235@siswa.ukm.edu.my, nazlia@ukm.edu.my