ANALISIS SENTIMEN DALAM BITCOIN TWEETS

LUM CHOI KIAN
DR WAN FARIZA FAUZI

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Bitcoin merupakan suatu "*cryptocurrency*" yang dicipta oleh Satoshi Nakamoto pada tahun 2009. Penggunaan Bitcoin semakin berleluasa sehingga melibatkan negara-negara seperti Nigeria , South Africa dan arus ini telah mengalir ke Malaysia sejak tahun 2016. Bank Negara Malaysia telah mengeluarkan kenyataan mengenai bitcoin yang tidak diiktiraf sebagai sah diperlakukan di Malaysia. Tetapi adakah keputusan kerajaan tersebut betul? Persoalan ini merupakan isu yang perlu diselidik sedangkan negara Venezuela mengunakan Bitcoins sebagai strategi demi memulihkan masalah ekonomi yang melanda negara mereka. Bitcoins bukan sahaja telah membawa impak positif tetapi juga mendatangkan pelbagai kesan buruk. Bitcoins mendatangkan banyak masalah sosial dan jenayah yang mampu mempengaruhi pembangunan negara**.** Oleh yang demikian, adalah agenda penting untuk mengetahui respon atau pandangan masyarakat terhadap Bitcoins. Dengan itu, analisis sentimen mengenai Bitcoins akan dijalankan untuk mengumpul data atau komen masyarakat supaya dapat membantu kita dalam membuat keputusan tentang hal yang berkaitan dengan Bitcoins. Projek tersebut telah menjalani perbandingan di antara 3 jenis model iaitu pengelas Naïve Bayes, Convolutional Neural Network dan Logistic Regression untuk menganalisis sentimen masyarakat terhadap Bitcoin. Hasil analisis tersebut telah memberi nilai data tentang respon masyarakat terhadap Bitcoins dan boleh dirujuk apabila memerlukan panduan dalam isu-isu Bitcoin terutamanya pelaburan Bitcoin.

1 PENGENALAN

Bitcoin merupakan suatu *cryptocurrency* yang dicipta oleh Satoshi Nakamoto pada tahun 2009, atau dalam bahasa yang lebih mudah, Bitcoin adalah matawang digital. Perbezaan Bitcoin

dengan virtual currency adalah Bitcoin boleh digunakan sebagai medium pertukaran untuk memiliki sesuatu barang dan perkhidmatan dalam dunia sebenar. Namun penggunaan Bitcoin mengandungi risiko yang tinggi serta membawa banyak impak negatif kepada sosial dan ekonomi negara. Hal ini kerana Bitcoin merupakan mata wang digital yang bersifat *decentralized* , atau kata lain Bitcoin tidak dapat dikawal oleh Bank Negara Malaysia seperti Ringgit Malaysia, Bitcoin tidak ada pengatur kewangan (*financial regulator*) seperti BNM.

Kegiatan Bitcoin semakin berleluasa sehingga melibatkan negara-negara seperti Nigeria , South Africa dan arus ini telah mengalir ke Malaysia sejak tahun 2016. Bank Negara Malaysia telah mengeluarkan kenyataan mengenai bitcoin yang tidak diiktiraf sebagai sah diperlakukan di Malaysia. Tetapi adakah keputusan tersebut betul? Walau bagaimanapun, masih ada sebahagian masyarakat yang terlibat dalam *cryptocurrency* Bitcoins tersebut. Apakah faktor yang menarik minat kelompok besar masyarakat tersebut sanggup melaburkan jumlah matawang yang besar dalam transaksi Bitcoins. Oleh yang demikian, saya telah menjalankan analisis sentimen untuk mengetahui respon dan pandangan masyarakat terhadap Bitcoin.

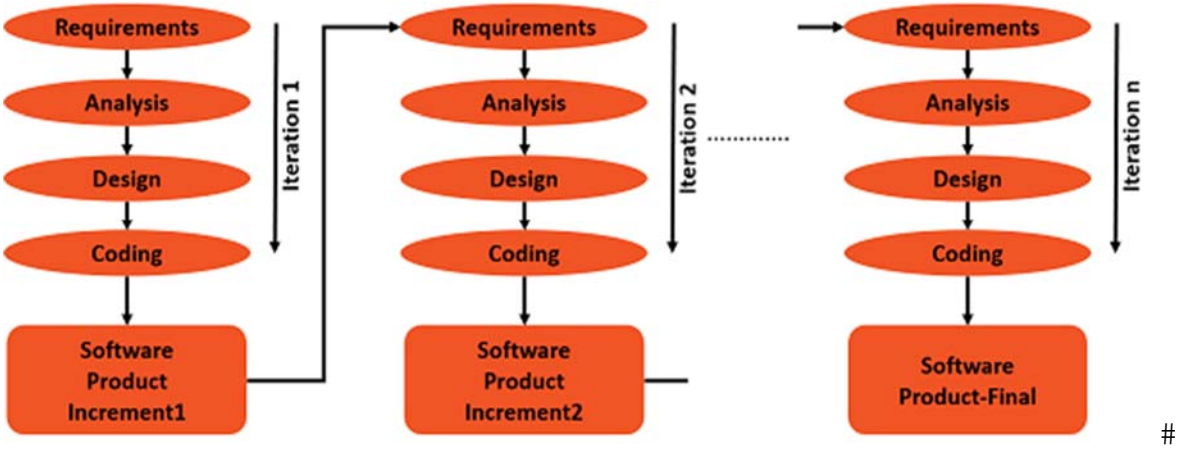
2 PENYELESAIAN MASALAH

Melalui kajian terhadap berita-berita tersebut, dapat disimpulkan bahawa Bitcoins memang banyak mempengaruhi kegiatan sosial dan ekonomi negara. Oleh yang demikian, adalah penting untuk mengetahui respon atau pandangan masyarakat terhadap Bitcoins. Dengan itu, analisis sentimen mengenai Bitcoins perlu dijalankan untuk pada data atau komen masyarakat mengenai Bitcoin supaya dapat membantu kita dalam membuat keputusan tentang hal yang berkaitan dengan Bitcoin. Dengan ini, masyrakat dan pihak berkuasa boleh menggunakan informasi tersebut untuk menilai semula tentang baik buruk yang dibawa oleh Bitcoins.

3 OBJEKTIF KAJIAN

Projek tersebut bertujuan mengumpul respon dan sentimen masyarakat terhadap Bitcoins dan mengkaji model pengelas yang paling sesuai untuk melatih serta mengira kebarangkalian ketepatan Bitcoin tweet bagi menghasilkan hasil data analisis yang mampu menjadi rujukan untuk masyarakat yang berminat. Misalnya, pelabur Bitcoins dan peniaga.

4 Methodologi



#

Rajah 1 Model Iteratif

(Anon, 2018)

Projek ini menggunakan model pengembangan Iteratif. Model Iterative merupakan model pengembangan sistem yang bersifat dinamis dalam artian setiap tahapan proses pengembangan sistem dapat diulang jika terdapat kekurangan atau kesalahan. Setiap tahapan pengembangan system dapat dikerjakan berupa ringkasan dan tidak lengkap, namun pada akhir pengembangan akan didapatkan sistem yang lengkap pada pengembangan system.

Iterative Development berarti menciptakan versi yang lebih fungsional dari sebuah sistem dalam siklus pembangunan pendek. Setiap versi ditinjau dengan pengguna untuk menghasilkan persyaratan untuk membuat versi berikutnya. Proses ini diulang sampai semua fungsionalitas telah dikembangkan. Panjang ideal iterasi adalah antara satu hari (yang lebih dekat dengan Metodologi Agile) dan tiga minggu. Setiap siklus pengembangan memberikan pengguna kesempatan untuk memberikan umpan balik,memperbaiki persyaratan, dan kemajuan melihat (dalam pertemuan sesi fokus grup). Hal ini akhirnya pembangunan berulang yang memecahkan masalah yang melekat dalam metodologi fleksibel dibuat pada 1970an.

Dalam model Iteratif ini, semua keperluan dibahagikan kepada pelbagai fasa. Semasa setiap iterasi, modul pembangunan menjalani fasa keperluan, reka bentuk, implementasi dan

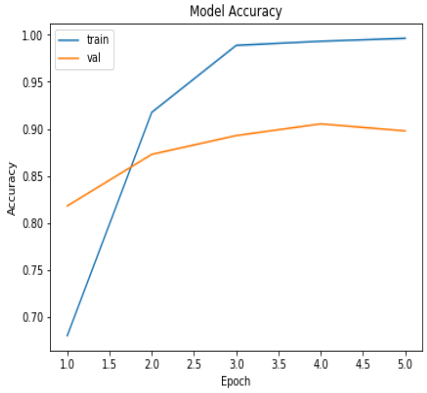
pengujian. Setiap modul seterusnya bertambah baik fungsi berbanding dengan keluaran sebelumnya. Proses ini berterusan sehingga sistem lengkap siap seperti kehendak.

Terdapat beberapa kebaikkan model Iteratif iaitu prototaip relatif lebih mudah dibangun dan tidak memerlukan waktu yang lama. Dengan prototaip, kesalahan & kelalaian dalam pengembangan dapat segera diketahui. Keputusan dapat diperoleh secara awal dan berkala. Pembangunan selari boleh dirancang. Tambahan pula, model Iteratif dapat mengurangkan kos untuk menukar skop / keperluan. Ujian dan debugging semasa setiap iterasi kecil adalah lebih mudah.

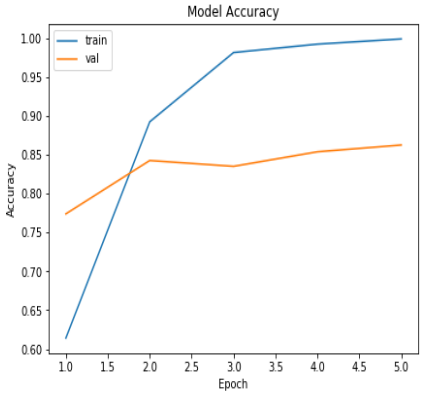
5 HASIL KAJIAN

5.1 Penetapan saiz set ujian CNN

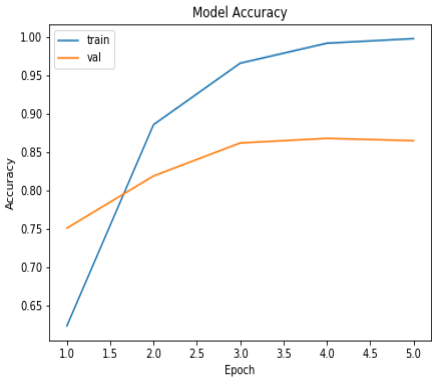
Saiz uji=0.2



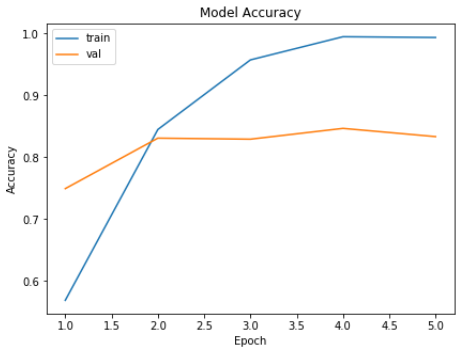
Saiz uji =0.4



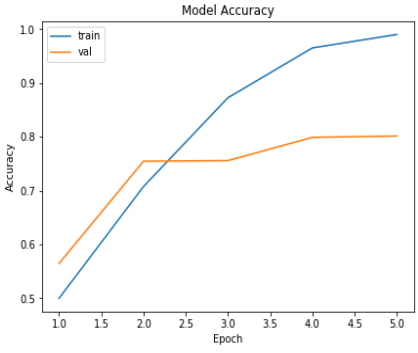
Saiz uji =0.5



Saiz uji =0.6



Saiz uji =0.8



#

Rajah 2 Model CNN Dengan Saiz Ujian Berbeza

Oleh itu, saiz uji yang paling sesuai untuk modal CNN adalah saiz 0.2. Hal ini kerana apabila saiz uji sama dengan 0.2, paling banyak data yang digunakan untuk melatih pengelas. Oleh yang demikian, ketepatan lebih tinggi berbanding dengan saiz uji yang lain.

5.2 Penetapan nilai *Epoch* dan *batch_size*

Satu epoch ialah apabila satu dataset selesai dimasukkan ke dalam *neural network* secara ulang alik sebanyak sekali.#Malah dalam scenerio realistik, satu set data terlalu besar untuk dimasuki sekaligus ke dalam pengelas. Oleh itu, satu set data dibahagikan lagi kepada beberapa kelompok yang lebih kecil.

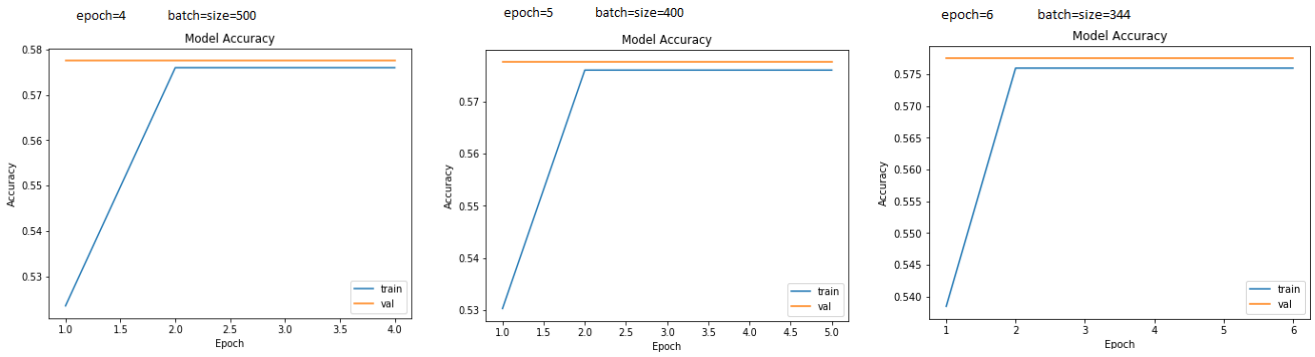
Jikalau nombor *Epoch* yang terlalu rendah, pengelas akan menjadi kurang tepat atas masalah *underfitting*. Sebaliknya, jika nombor *Epoch* yang terlalu tinggi, pengelas dilatih secara keterlaluan dan membawa kepada *overfitting*.

Batch_size adalah jumlah data latihan yang wujud dalam satu *batch* . Bilangan batch dan *batch_size* dua perkara yang berbeza.

Sebagai contoh, katakan kita mempunyai sebanyak 2000 data latihan. Data latihan 2000 dibahagi kepada beberapa kelompok *batch*, di mana saiz batch adalah 500 dan Iterasi adalah 4, untuk melengkapi satu epoch.

Berdasarkan Rajah 5.8 saiz uji 0.2 di atas, ketepatan pengelas meningkat secara drastik dengan peningkatan Epoch sehingga 3.0. Graf terus meningkat dan mencapai kestabilan apabila medekati Epoch 5.0. Oleh itu, jangkaan boleh dibuat bahawa pengelas CNN mencapai prestasi terbaik apabila epoch sama dengan 5.0.

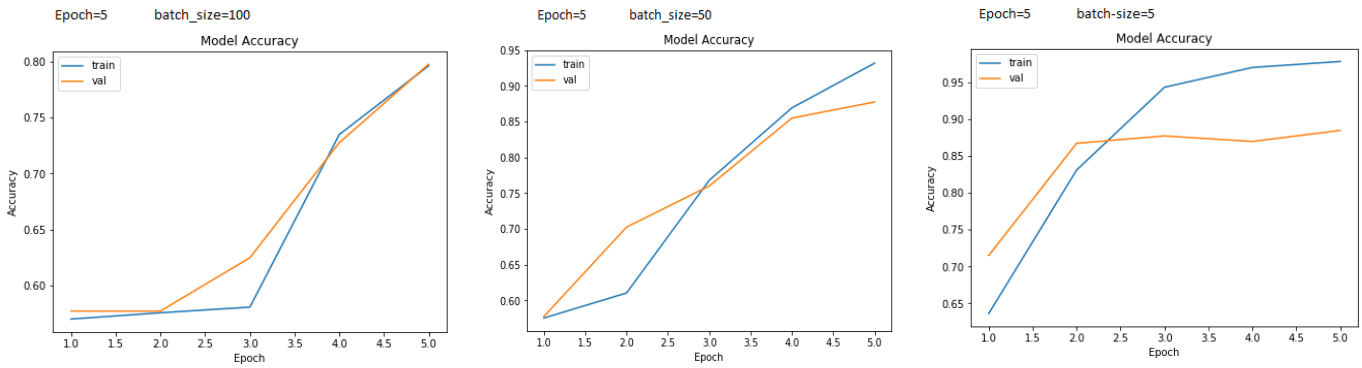
Jangkaan tersebut boleh terus dibuktikan menguji hasil CNN dengan “Epoch”#(Caudill al et., 1994) yang berbeza.



Rajah 3 Model CNN Dengan Epoch Berbeza

Berdasarkan graf diatas, hasil ketepatan dengan diuji dengan epoch yang berbeza iaitu 4,5,6. Graf menunjukan sedikit peningkatan apabila menambah epoch 4 kepada 5 dan ketepatan hampir sama bagi epoch 6. Dengan itu, Epoch 5 merupakan nilai yang bagi sesuai bagi pengelas CNN.

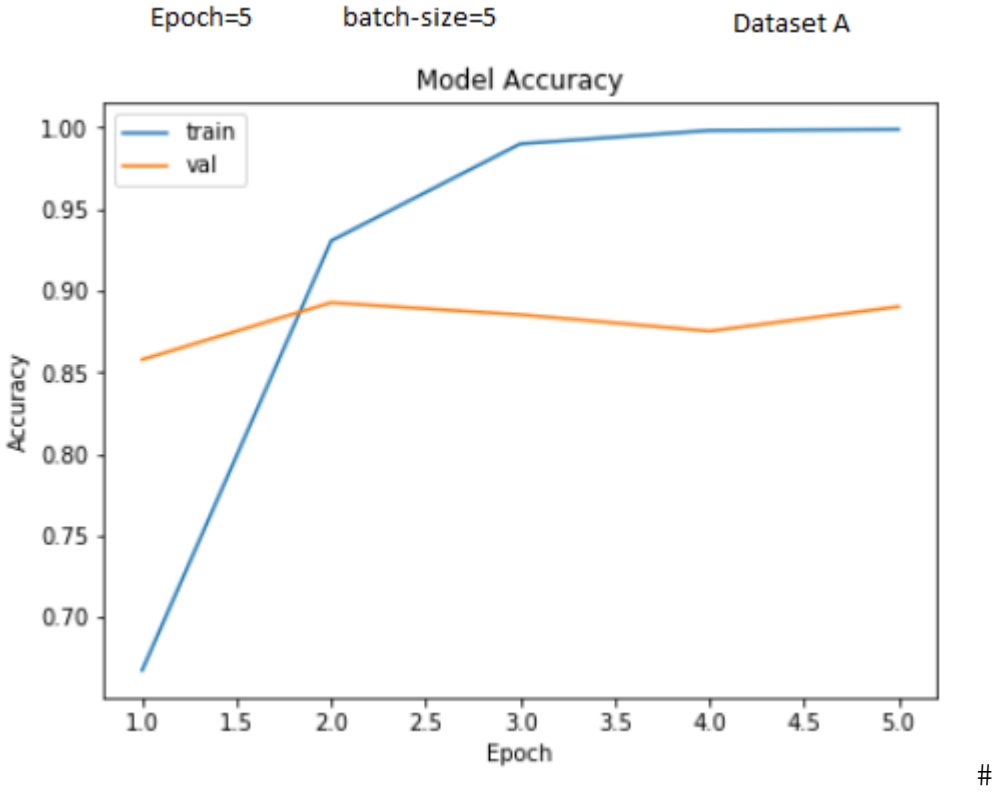
Walaupun CNN mempunyai prestasi yang baik apabila epoch bersamaan dengan 5, maka pada tahap tersebut ketepatannya masih rendah iaitu menghampiri 0.58. Justeru itu, "*batch_size*" perlu dilaraskan untuk meningkatkan ketepatan(Sharma, S. A. G. A. R., 2017).



Rajah 4 Model CNN Dengan Saiz Batch Berbeza

Graf-graf di atas menunjukkan pengaruh "*batch_size*" (Caudill & Butler, 1994) terhadap ketepatan pengelas. Dengan penurunan saiz batch, terdapat peningkatan ketara pada ketepatan dan ia mencapai maksima apabila saiz batch bersamaan dengan 5.

5.3 Hasil *Convolutional Neural Network*



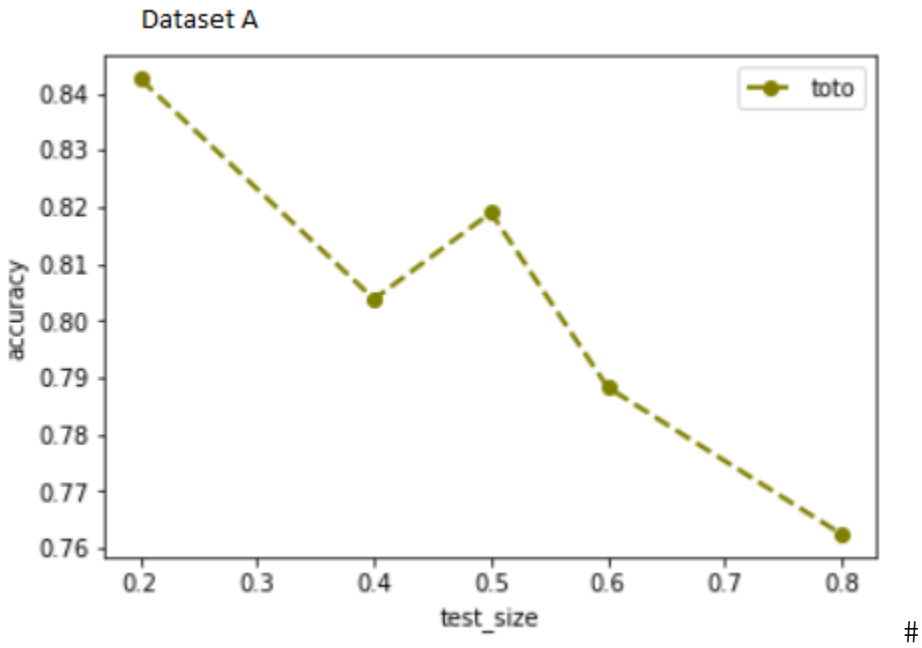
#

Rajah 5 Ketepatan Model CNN

Ketepatan terus ditingkatkan dengan CNN sehingga menghampiri ketepatan maxima. CNN mencapai ketepatan terbaik berbandingan dengan kedua-dua model yang lain mungkin kerana ketinggian varians set data projek tersebut. Ciri model CNN yang mampu mengawal size penapis mampu mendapat corak terperinci set data secara lebih menyeluruh. Dengan itu, CNN merupakan pengelas yang paling sesuai bagi set data projek tersebut.

5.4 Hasil *Naïve Bayes*

Berdasarkan rajah di bawah, ketepatan yang diperoleh dengan *Naïve Bayes* tidak begitu memuaskan. Ketepatan yang paling tinggi mencapai 0.84. Seterusnya model *Logistic Regression* digunakan dengan harapan meningkatkan ketepatan.

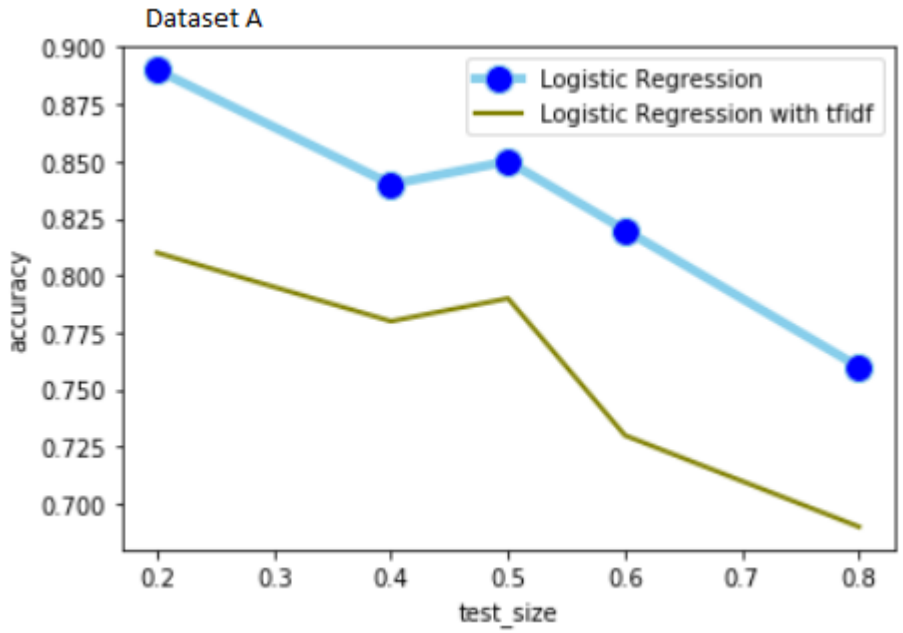


#

Rajah 6 Ketepatan Model Naïve Bayes

5.5 Hasil *Logistic Regression*

Berdasarkan graf di atas, model *Logistic Regression* berjaya meningkatkan ketepatan sebanyak 0.049 dengan ketepatan tertinggi sehingga 0.889. Logistic Regression dengan tfidf menunjukkan ketepatan yang lebih rendah model *Logistic Regression.* Hal ini mungkin kerana ciri tfidf yang mengambil berat tentang hubungan di antara perkataan tetapi set data yang digunakan mempunyai variasi yang tinggi. Oleh yang demikian, tfidf tidak mendapat ketepatan yang memuaskan.



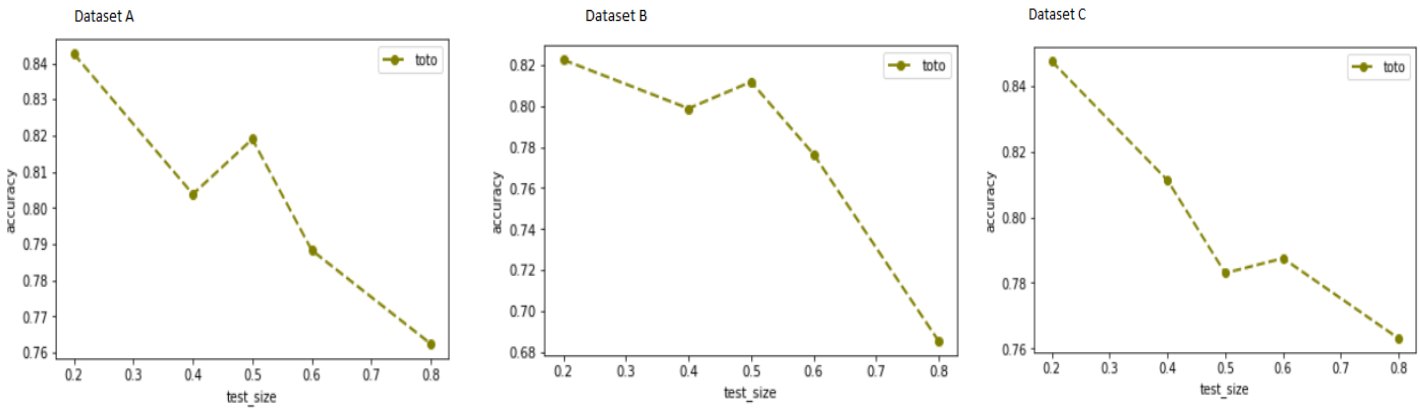
#

Rajah 7 Ketepatan Model Logistic Regression

5.6 Perbandingan model dengan dataset berbeza

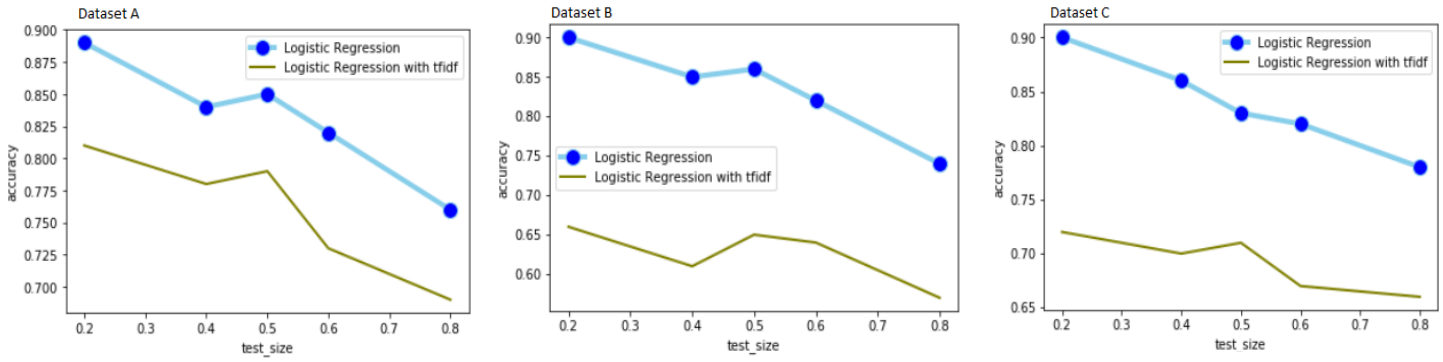
Projek tersebut akan mempuyai 3 set data yang merangkumi tempoh masa yang berbeza iaitu Set A adalah set data pada awalan tahun 2018, Set B adalah set data akhiran tahun 2018, manakala Set C pula adalah set data awalan tahun 2019. Set-set data tersebut masih baharu dan sesuai untuk tugasan analisis. Selain itu, Set B dan Set C adalah set data standard emas.

Rajah dibawah akan menunjukkan pengaruh data tersedia(Set A) dengen set data standard emas(Set B & Set C) terhadap ketepatan pelbagai modal:



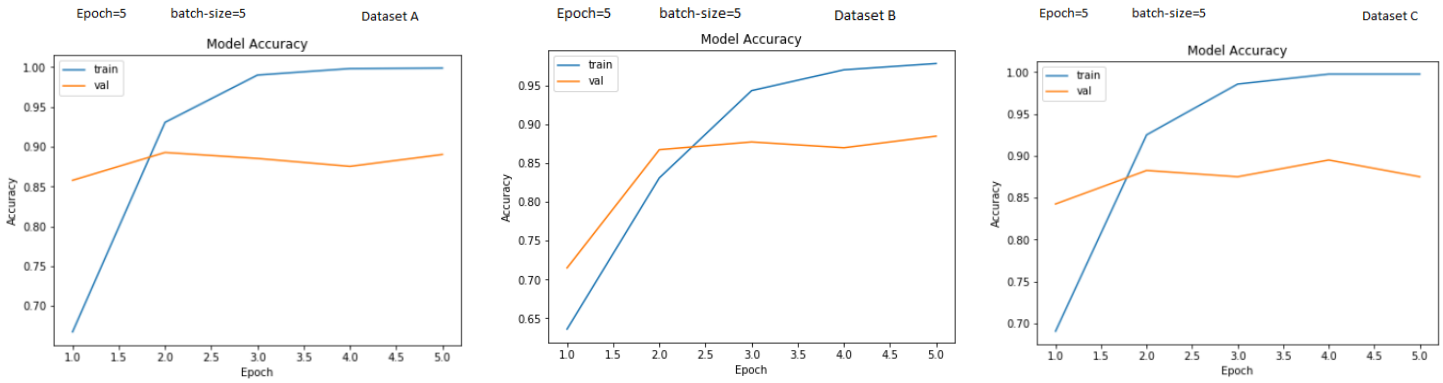
Rajah 8 Perbandingan NB SetA Dengan Data Standard Emas

Berdasarkan rajah di atas, bagi modal Naïve Bayes set data C mempunyai ketepatan yang lebih tinggi daripada set data A dan B. Hal ini mungkin kerana *feature* yang digunakan seperti word frekuensi bagi sesuai untuk set C. Dengan kata lain, pengelas set C dapat memperoleh *feature* yang lebih tepat berbanding dengan set A dan B. Maka, ketepatan tersebut mungkin akan menurun apabila parameter yang sama diimplimentasi keatas data lain seperti A dan B. Penurunan ketepatan adalah disebabkan oleh masalah underfitting iaitu kegagalan pengelas untuk mengesan ciri-ciri set data secara tepat.



Rajah 9 Perbandingan LR SetA Dengan Data Standard Emas

Berdasarkan rajah di atas , bagi modal Logistic Regression set data B dan C mempunyai ketepatan yang lebih tinggi daripada set data A. Hal ini mungkin kerana parameter yang ditetapkan seperti jumlah ngram_range, c-value lebih sesuai bagi set B dan C. Dengan kata lain, pengelas set B dan C dapat memperoleh *feature* yang lebih tepat berbanding dengan set A dan B. Pengelas tersebut mungkin akan menhadapi masalah overfitting apabila pengelas tersebut digunakan untuk mengklasifikasikan set data yang baharu.



Rajah 10 Perbandingan CNN SetA Dengan Data Standard Emas

Berdasarkan rajah di atas , bagi modal *Convolutional Neural Network*. Ketiga-tiga set data memperoleh hasil ketepatan yang amat memuaskan. Hal ini mungkin kerana parameter yang ditetapkan seperti jumlah penapis, epoch dan batch saiz lebih sesuai bagi semua set. Dengan kata lain, pengelas dapat mengesan *feature* dengan sangat tepat. Namun, pengelas tersebut mungkin akan menghadapi masalah overfitting apabila pengelas tersebut digunakan untuk mengklasifikasikan set data yang baharu.

6 KESIMPULAN

Kesimpulannya, pembangunan analisis sentimen mengenai Bitcoins Tweets telah berjaya dibangunkan. Dalam kajian tersebut model Convolutional Neura Network dipilih kerana ia menghasilkan ketepatan yang paling tinggi berbanding dengan modal Naïve Bayes dan Logistik Regression. Walau bagaimanapun, analisis yang lebih mendalam seperti pertambahan teknik-teknik lain seperti TF-IDF atau N-gram boleh dijalankan bagi meningkatkan lagi ketepatan kajian.Diharapkan dengan analisis tersebut, ia boleh menjadi perintis rujukan untuk pelabur Bitcoins dan juga pihak kerajaan. Analisis tersebut dapat memberi informasi berguna tentang pandangan masyarakat terhadap Bitcoins dengan harapan analisis tersebut secara tidak langsung dapat membawa impak positif kepada negara.

7 RUJUKAN

Sharma, S. A. G. A. R. (2017). Epoch vs batch size vs iterations. *Towards Data Science*, *23*.

Caudill, M., & Butler, C. (1994). *Understanding neural networks: computer explorations: a workbook in two volumes with software for the macintosh and pc compatibles*. MIT press.