PENGECAMAN BAHASA ASAL MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN

Ng Kar Lun

Dr. Wan Fariza Fauzi

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Penipuan siber telah wujud sejak bermulanya penggunaan teknologi yang bertujuan untuk berkomunikasi di masyarakat kita. Teknologi yang berkembang pesat ini telah mewujudkan lebih laman web atau aplikasi yang bertujuan untuk berkomunikasi dan telah menyebabkan penipuan siber semakin meningkat di Malaysia pada beberapa tahun ini. Kebanyakan penipu siber adalah dari Nigeria dan mereka mencari mangsa melalui laman web sosial. E-mel yang dihantar oleh penipu biasanya dalam Bahasa Inggeris dan penipu biasanya berpura sebagai bangsa Eropah. Rakyat Malaysia mudah jatuh ke perangkap penipu adalah kekurangan pengetahuan mengenai perbezaan linguistik yang digunakan di negara lain. Untuk mengurangkan peluang rakyat Malaysia jatuh ke perangkap penipu, aplikasi pengenalan bahasa yang dibangunkan ini dapat mengira kebarangkalian mesej yang diterima oleh individu adalah dari bangsa mana. Dalam kes yang banyak penipu siber adalah dari Nigeria, aplikasi ini akan fokus kepada mengkaji linguistik orang Nigeria dengan mencari keunikan aksen mereka. Dalam projek ini, kami akan menggunakan ciri-ciri unigram dan pengagihan frekuensi untuk melatih dan menguji pengelas yang dipilih untuk digunakan dalam projek ini.

1 PENGENALAN

Teknologi yang canggih pada masa kini banyak memudahkan kehidupan harian kita dari pelbagai aspek dan teknologi yang paling berkembang pesat dalam beberapa tahun ini merupakan internet. Selain menggunakan internet untuk tujuan hiburan, ramai orang menggunakan internet untuk tujuan berkomunikasi. Dengan adanya e-mel, *WhatsApp*, *Facebook* dan laman web atau aplikasi media sosial yang lain telah banyak memudahkan komunikasi antara manusia. Disebalik kemudahan yang disumbangkan oleh kepesatan teknologi internet masa kini, terselit seribu macam bahaya dan risiko yang perlu ditanggung oleh para pengguna dan di antaranya termasuklah aktiviti penipuan siber (*online scamming*). Hal sedemikian terjadi disebabkan keperluan pengguna untuk mendedahkan maklumat peribadi seperti nama, e-mel dan nombor telefon sebelum dapat menggunakan perkhidmatan

seperti laman sesawang berdaftar dan aplikasi media sosial. Inilah yang akan menyebabkan kebocoran maklumat peribadi kita kerana sesetengah laman web dan aplikasi media sosial akan menjual maklumat peribadi pengguna mereka untuk mendapatkan duit dan inilah cara ramai penipu siber dapat mencari mangsa dan menghantar e-mel atau mesej kepada mangsa mereka.

Penyelidikan yang dijalankan oleh *Telenor Group* menunjukkan Malaysia merupakan negara yang paling terdedah kepada penipuan siber berbanding dengan negara jiran iaitu 46% daripada responden kajian selidik merupakan mangsa penipuan siber ("Malaysia is the most vulnerable country to internet scams in this region - Business News | The Star Online," 2016). Kelonggaran visa pelajar dan sistem perbankan berteknologi tinggi menyebabkan Malaysia menjadi hab global untuk penipuan internet. Pada tahun 2013, Malaysia diberikan kedudukan keenam oleh SOPHOS (pembangun keselamatan teknologi maklumat Amerika Syarikat) dari segi jenayah siber (Charlie Campbell, 2014).

Antara penipuan siber yang sering berlaku ialah: e-mel palsu bertujuan untuk mencuri maklumat peribadi dalam talian (*phishing*), penipuan Nigeria atau 419 (*Nigerian or* 419 *scams*), penipuan cinta siber (*romance scams*), penipuan pekerjaan (*employment scams*), dan sebagainya. Fokus projek ini adalah pada penipuan Nigeria dan penipuan cinta siber.

Penipuan sebegini menggunakan Internet sebagai medium untuk berkomunikasi dan melibatkan penulisan text. E-mel dan laman media sosial yang sering digunakan untuk mewujudkan hubungan. E-mel mempunyai fungsi untuk mengenal pasti spam dan memindahnya ke *junk*, tetapi melalui laman sosial media contohnya laman web jodoh penipu boleh mendapatkan maklumat mangsa dan menghantar mesej kepada mereka. Dalam kes penipuan cinta siber, penipu akan menyamar sebagai seorang professional yang berbangsa Eropah. Mereka cuba menambatkan hati mangsa dan kemudian diceritakan pelbagai masalah yang memerlukan bantuan kewangan.

Oleh demikian, projek ini yang melibatkan pembangunan aplikasi linguistik untuk pengecaman bahasa yang digunakan dalam teks mesej atau e-mel dicadangkan. Dengan aplikasi ini, kredibiliti teks tersebut boleh dikenalpasti dan secara tidak langsung dapat membantu rakyat Malaysia supaya tidak mudah untuk jatuh ke perangkap penipuan ini.

2 PENYATAAN MASALAH

Penipuan siber telah wujud sejak bermulanya penggunaan teknologi yang bertujuan untuk berkomunikasi di masyarakat kita. Teknologi yang berkembang pesat ini telah mewujudkan lebih laman media sosial dan aplikasi yang bertujuan untuk berkomunikasi, contohnya E-mel, *WhatsApp*, *Facebook*, *Instagram*, laman-laman web jodoh seperti *www.match.com*. Perkembangan ini telah menyebabkan penipuan siber semakin meningkat di seluruh dunia termasuk Malaysia.

Kebanyakan penipu siber adalah dari Nigeria sehingga wujudnya satu kategori kes penipuan Nigeria atau 419 (Hiss, 2015). Kebanyakan kes penipuan cinta juga melibatkan rangkaian penjenayah dari Nigeria (Hamsi, Bahry, Tobi, & Masrom, 2015). Menurut statistik dalam *SCAMWATCH*, bilangan kes yang dilapor di Australia pada tahun 2017 ialah 998 kes dan jumlah kehilangan harta kira-kira 1.5 juta *U.S. Dollar* ("Scam statistics | Scamwatch," 2017). Di Malaysia, dari tahun 2015 hingga 2016, *CyberSecurity* Malaysia (CSM) mencatatkan sekurang-kurangnya 3257 kes jenayah siber ("Kes Penipuan Dalam Talian Meningkat - Madius - Berita Jenayah | mStar," 2017). Mereka mencari mangsa melalui laman web jodoh seperti yang dinyatakan di atas, sasaran mereka biasanya ialah wanita dalam lingkungan pertengahan umur dan berasa kesepian. Selain penipuan percintaan, terdapat penipuan lain iaitu penipuan pelaburan atau perniagaan yang mana penipu akan menyatakan bahawa mereka memerlukan pelaburan daripada mangsa dan sebagai balasan akan memperoleh ganjaran yang lumayan. Mereka akan menggunakan profil palsu dan akan berpura sebagai bangsa Eropah semasa mendekati mangsa, biasanya melalui E-mel. Melalui maklumat yang diperolehi dari laman web sosial yang ada, e-mel yang dihantar oleh penipu tidak akan dianggap sebagai spam untuk dipindah kepada *junk*.

Rakyat Malaysia mudah jatuh ke perangkap penipuan siber ini. Dengan kebolehan untuk mengenalpasti perbezaan linguistik antara Bahasa Inggeris *native* (contoh: *British English*, *American English*) dan *non-native* iaitu kaum yang menggunakan Bahasa Inggeris sebagai bahasa kedua (contoh: *Malaysian English, Nigerian English*), rakyat Malaysia dapat megecam penipuan yang terkandung dalam teks E-mel.

3 OBJEKTIF KAJIAN

Projek ini adalah untuk membina aplikasi yang dapat megecam bahasa dengan pengiraan kebarangkalian suatu teks berdasarkan ciri-ciri linguistik unik yang terdapat dalam penggunaan Bahasa Inggeris sebagai bahasa kedua oleh bangsa Nigeria dan Malaysia. Bagi meningkatkan ketepatan aplikasi ini, objektif berikut perlu dicapai:

i. Mengumpul data tentang linguistik orang Nigeria dan menganalisi keunikan dialek mereka.
ii. Mengenal pasti teknik pembelajaran mesin yang paling sesuai digunakan untuk pengecaman bahasa.
iii. Menguji dengan data baru untuk memperoleh teknik pembelajaran mesin yang paling sesuai.

4 METOD KAJIAN

Metodologi yang digunakan dalam projek ini adalah kitaran pembangunan *Agile* (*Agile Development Cycle*). Kelebihan terbesar *Agile* berbanding dengan metodologi lain ialah fleksibilitinya yang mana ia boleh dipinda untuk memuaskan keperluan projek. Dengan metodologi *Agile*, produk akhir yang dihasilkan mempunyai kecacatan yang berkurangan. Berikut adalah fasa-fasa yang ada dalam kitaran pembangunan *Agile*:

4.1 Fasa Keperluan (*Requirements*)

Fasa ini adalah sebelum bermula untuk melaksanakan projek. Dalam fasa ini, kita perlu mengenal pasti apakah aplikasi yang perlu dibangun dan mengumpul data tentang keperluan minimum untuk membangunkan aplikasi tersebut.

4.2 Fasa Perancangan (Plan)

Fasa ini adalah semasa memulakan projek, kita perlu mengenal pasti objektif projek dan cadangan penyelesaian bagi penyataaan masalah projek supaya pelaksanaan projek akan mengikut apa yang ditetapkan. Pengumpulan data dan analisi data juga akan dilaksanakan di fasa ini.

4.3 Fasa Reka Bentuk (Design)

Fasa ini akan fokus terhadap bagaimana aplikasi ini dapat dilaksanakan. Antara muka akan direka bentuk kalau aplikasi projek mengandungi sebarang antara muka.

4.4 Fasa Pembangunan (Develop)

Fasa ini adalah untuk membangunkan sistem utama aplikasi projek yang telah ditetapkan dalam fasa sebelum ini. Pengekodan atas fungsi yang telah direka bentuk dalam antara muka akan dilakukan dan juga pengujian untuk memastikan aplikasi dapat berfungsi dengan betul. Selepas siap pengekodan atas semua fungsi, aplikasi akan diuji oleh beberapa pengguna untuk mendapatkan komen untuk menambahbaikkan aplikasi

4.5 Fasa Pelepasan (Release)

Fasa ini ialah fasa yang mana produk akhir sudah siap dihasilkan dan akan dilepaskan kepada pengguna untuk digunakan dalam kehidupan harian mereka.

4.6 Fasa Penjejakan & Pemantauan (Track & Monitor)

Fasa ini merupakan fasa yang tempohnya panjang sebab produk yang sudah dilepaskan akan terus diawasi untuk memastikan produk dapat berfungsi dengan lancar dan juga untuk mencari penyelesaian apabila mempunyai kecacatan yang tidak dapat dikesan dalam fasa-fasa sebelum ini dengan secepat mungkin. Selain itu, komen daripada semua pengguna akan dikumpulkan bagi kerja-kerja menaik taraf produk dapat dilakukan dengan lebih mesra pengguna.

6 HASIL KAJIAN

Bab ini adalah untuk menerangkan fasa pembangunan bagi aplikasi linguistic projek ini. Fasa reka bentuk yang telah diterangkan dalam bab sebelum ini akan disebutkan dalam bab ini. Ini kerana fasa pembangunan aplikasi adalah selepas fasa reka bentuk aplikasi dan kedua-dua fasa berkait rapat. Bagi memastikan aplikasi linguistik projek ini dapat dijalankan dengan lancar dalam semasa fasa pelepasan, proses pengujian aplikasi juga akan diterangkan dalam bab ini.

6.1 Melatih Data Dan Menguji Data

Set data yang telah siap diubah suai adalah bersedia untuk dilatih dan diuji oleh pengelas (*classifier*) yang telah dipilih. Langkah-langkah tersebut diperlukan untuk melatih dan menguji set data. Pertama sekali, adalah untuk membuka dan membaca setiap fail teks yang terdapat dalam kedua-dua folder tersebut dan *shuffle* rawak semua fail teks yang terdapat dalam kedua-dua folder tersebut untuk mengelakkan bias. Langkah seterusnya adalah untuk mendapatkan pengagihan frekuensi (*frequency distribution*) bagi setiap perkataan yang terdapat dalam semua fail teks dalam kedua-dua folder tersebut dan selepas itu adalah untuk mengeluarkan 1000 perkataan yang teratas yang mana pengagihan frekuensi bagi kebanyakan perkataan adalah lebih daripada 100 bagi tujuan melatih dan menguji set data. Sebelum set data dapat dilatih dan diuji, satu langkah perlu dilakukan adalah membahagikan set data yang dalam urutan rawak kepada set melatih dan set menguji, bagi projek ini adalah untuk membahagikan 2/3 daripada jumlah set data kepada set melatih dan 1/3 daripada jumlah set data kepada set menguji. Proses melatih dan menguji dapat dijalankan setelah set data telah dibahagikan dan ketepatan bagi setiap pengelas dalam proses tersebut juga akan ditunjukkan. Jadual berikut menunjukkan ketepatan bagi setiap pengelas.

Rajah 1 Ketepatan pengelas yang digunakan

Menurut jadual di atas, setiap pengelas yang telah dilatih dan diuji dengan set data yang diberi dengan menggunakan pengagihan frekuensi bagi setiap perkataan dalam ciri *unigram* telah menunjukkan ketepatan yang agak tinggi. Oleh itu, kami menjangka bahawa setiap pengelas tersebut dapat mengklasifikasikan dokumen yang baru dengan tepat. Kami juga

menggunakan pengagihan frekuensi bagi setiap perkataan dalam ciri *bigram* dan berikut adalah keputusan peringkat melatih dan menguji bagi setiap pengelas.

```
'''
 RESTART: C:\Users\gonemiss\Desktop\New folder\NLP\Normal Set\1Naive Bayes.py
Naive Bayes Algo accuracy: 49.705882352941174
Most Informative Features
          ('give', 'god') = False          NonNig : Nigeri =      1.0 : 1.0
   ('ok', 'incidentally') = False          NonNig : Nigeri =      1.0 : 1.0
          ('stop', 'aha') = False          NonNig : Nigeri =      1.0 : 1.0
       ('people', 'like') = False          NonNig : Nigeri =      1.0 : 1.0
          ('sin', 'ones') = False          NonNig : Nigeri =      1.0 : 1.0
    ('niile', 'naenyere') = False          NonNig : Nigeri =      1.0 : 1.0
         ('love', 'also') = False          NonNig : Nigeri =      1.0 : 1.0
       ('want', 'poland') = False          NonNig : Nigeri =      1.0 : 1.0
        ('could', 'part') = False          NonNig : Nigeri =      1.0 : 1.0
          ('yab', 'yab') = False          NonNig : Nigeri =      1.0 : 1.0
       ('merely', 'go') = False          NonNig : Nigeri =      1.0 : 1.0
    ('profession', 'get') = False          NonNig : Nigeri =      1.0 : 1.0
         ('itll', 'one') = False          NonNig : Nigeri =      1.0 : 1.0
       ('jooka', 'thank') = False          NonNig : Nigeri =      1.0 : 1.0
          ('god', 'go') = False          NonNig : Nigeri =      1.0 : 1.0
MultiNB_classifier accuracy: 49.705882352941174
SVC_classifier accuracy: 49.705882352941174
LinearSVC_classifier accuracy: 49.705882352941174
'''
```

Rajah 2 Ketepatan pengelas bagi ciri *bigram*

Daripada jadual tersebut, ketepatan bagi setiap pengelas adalah agak rendah dan ciri yang ditunjukkan oleh *Naïve Bayes Classifier* adalah dalam nisbah 1:1 yang mana ciri tersebut adalah ciri yang tidak berguna. Keputusan peringkat melatih dan menguji pengelas adalah sama semasa menggunakan pengagihan frekuensi bagi setiap perkataan dalam ciri *trigram*, *quadgram* dan *pentagram*.

6.2 Menguji Pengelas Menggunakan Data Baru

Setelah proses melatih dan menguji data sudah selesai dijalankan, pengelas yang telah dilatih boleh digunakan untuk mengklasifikasi dokumen yang baru untuk mencari pengelas yang mana adalah paling sesuai untuk projek ini. Sebelum dokumen yang baru dapat diuji, langkah yang diambil semasa penyediaan set data perlu dilakukan kepada dokumen baru ini iaitu menukarkan semua perkataan dalam teks kepada huruf kecil, membuangkan tanda baca dan nombor dalam teks serta membuangkan *stopwords* yang terdapat dalam teks. Pengujian bagi dokumen baru dapat dijalankan setelah dokumen baru ini telah siap diubah suai. Bagi menguji ketepatan pengelas yang telah dilatih dalam projek ini, 4 dokumen baru yang mana 2 dokumen adalah dalam kategori "*Nigerian*" dan 2 yang lain adalah dalam kategori "*NonNigerian*" telah

disediakan. Setiap dokumen akan diuji 5 kali bagi menguji konsistensi pengelas yang telah digunakan. Berikut adalah keputusan pengujian dokumen "*Nigerian*" dan "*NonNigerian*".

	Naïve Bayes	**Multinomial**	**SVC**	**Linear SVC**
Nig1	0/5	4/5	5/5	5/5
NonNig1	5/5	4/5	0/5	0/5
Nig2	0/5	3/5	5/5	5/5
NonNig2	3/5	3/5	0/5	0/5

Rajah 3 Keputusan keseluruhan pengujian

Keputusan keseluruhan pengujian di atas menunjukkan, daripada 5 kali pengujian bagi setiap dokumen yang baru, hanya *Multinomial Naïve Bayes Classifier* dapat mengklasifikasikan dokumen yang baru kepada kategori yang betul lebih banyak kali berbanding pengelas yang lain. Selain itu, daripada jadual di atas juga dapat dilihat bahawa *Naïve Bayes Classifier* mempunyai bias terhadap kategori "*NonNigerian*" dan *SVC*, *Linear SVC classifier* mempunyai bias terhadap kategori "*Nigerian*". Oleh itu, *Multinomial Naïve Bayes Classifier* adalah lebih sesuai digunakan dalam projek ini berbanding dengan pengelas yang lain bagi pengagihan frekuensi bagi setiap perkataan dalam ciri *unigram*.

6 KESIMPULAN

Kesimpulan bagi keseluruhan perancangan projek ini akan dibincangkan dalam Bab ini. Keseluruhan projek ini telah dibangunkan dalam tempoh masa yang ditetapkan dan dapat memenuhi objektif yang dinyatakan dalam projek ini. Selain membincangkan kekangan aplikasi yang telah dibangunkan dalam projek ini, Bab ini turut membincangkan cadangan untuk menambahbaikkan aplikasi ini berdasarkan ilmu pengetahuan yang diperoleh melalui kajian keseluruhan projek ini.

7 RUJUKAN

Berikut adalah rujukan yang digunakan:

Charlie Campbell. (2014). Malaysia Is Becoming a Global Hub For Internet Scams | Time. Retrieved December 10, 2017, from http://time.com/2968765/malaysia-is-becoming-a-global-hub-for-internet-scams-preying-on-the-lovelorn/

Hamsi, A. S., Bahry, F. D. S., Tobi, S. N. M., & Masrom, M. (2015). Cybercrime over Internet Love Scams in Malaysia: A Discussion on the Theoretical Perspectives, Connecting Factors and Keys to the Problem. *Journal of Management Research*, *7*(2), 169. https://doi.org/10.5296/jmr.v7i2.6938

Hiss, F. (2015). Fraud and Fairy Tales : Storytelling and Linguistic Indexicals in Scam E-mails. *International Journal of Literary Linguistics*, *4*(1), 1–23.

Kes Penipuan Dalam Talian Meningkat - Madius - Berita Jenayah | mStar. (2017). Retrieved December 11, 2017, from http://www.mstar.com.my/berita/berita-jenayah/2017/02/07/kes-jenayah-siber/

Malaysia is the most vulnerable country to internet scams in this region - Business News | The Star Online. (2016). Retrieved December 10, 2017, from https://www.thestar.com.my/business/business-news/2016/03/11/malaysia-is-the-most-vulnerable-country-to-internet-scams-in-this-region/

Scam statistics | Scamwatch. (2017). Retrieved December 11, 2017, from https://www.scamwatch.gov.au/about-scamwatch/scam-statistics?scamid=6&date=2017