Performance Evaluation of Analytics Models for Trends Analysis of News

Lau Yee Heng*, Rajasvaran Logeswaran and Booma Poolan Marikannan

School of Computing Asia Pacific University of Technology and Innovation Technology Park Malaysia, Bukit Jalil, 57000 Kuala Lumpur, Malaysia

Corresponding author e-mail: auyh87@gmail.com,

Abstract. Microblogging services, especially Twitter, allow the user to share their most recent thoughts, feelings or news freely and almost immediately. Hence, the number of news tweets generated by the news media is increasing exponentially. Mining the valuable data from the large volume of tweets can help increase the revenue of organisations by allowing them to engage with the public faster and better by responding to the latest topics of interest. In this work, mining the hot keywords and being able in classifying the news tweets, trending topic and keywords in the news tweets. Both supervised and unsupervised machine learning models are used. Several machine learning algorithms are being used to compare the accuracy in classifying the tweets.

Index Terms. Classification algorithms, Recurrent neural networks, Supervised learning, Text mining, Unsupervised learning, Web mining

1. Introduction

Twitter's fast pace has made it ‘the place’ where news breaks happen, attracting millions of people to keep up with the latest news in the world [1]. According to [2], there are 2.4 million twitter users in Malaysia, therefore, the number of Malaysian Twitter users that engage with the breaking news in Twitter is huge. In addition, according to [3], 70% percent of Malaysian read the news online.

The aim of this work is to analyze the news tweets to identify current hot keywords of each topic and perform news tweets classification. To achieve this objective, both supervised and unsupervised learning algorithms are used. Supervised machine learning model is used to classify the tweets while an unsupervised learning model will be used to relate a large amount of vectorized text to the limit number of unassigned topics. The supervised learning algorithms are logistic regression, support vector machine (SVM), artificial neural network (ANN), recurrent neural network (RNN) and long short-term memory (LSTM), whilst unsupervised learning algorithm is the latent dirichlet allocation (LDA).

Logistic regression is named after the logistic function used at the core of the method [4]. It is the old gold standard in performing classification. SVM are useful in performing data classification [5]. They identify a hyperplane in the N-dimensional space for N features to distinctly classify the data points [6]. ANN are mathematical models that try to simulate the structure and functionalities of biological neural networks [7]. Back propagation is used to update the weight for each iteration of training. RNN is a class of neural networks that is capable of solving the problem in modeling sequence data such as
time series or natural language [8]. With one or more feedback connections, the RNN enables the activations to flow around in a loop; thus, RNN is able to perform temporal processing and learn sequences. It has an additional node to loop back the internal state to process the input.

LSTM is an enhanced version of the RNN to overcome the issue of vanishing gradient problem (i.e. the decay of the loss function over the time) [9]. This is because RNN cannot store the memory for an extended period of time due to the context layer constantly updating the weight throughout the training phase [9]. The illustration of the LSTM’s architecture is shown in Fig. 1 [9].

![LSTM's architecture](image)

LDA is good at finding word-level topics, hence it is useful for clustering documents based on the topic [10]. It is a three-level hierarchical Bayesian model, that is used to identify the topic of the document by reducing the dimension of the document and calculate the probability distribution of the frequency of the words without the class labels [10]. LDA can be represented by:

$$p(\theta|z,w|\alpha,\beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(w_n|z_n,\beta)$$

(1)

2. Materials and Methods

2.1. Tweets

Twitter is a popular social networking and microblogging site that allows users to describe their current status within 140-character messages known as tweets [11]. Due to the properties of real-time sharing, Twitter has become the source of text information as more and more users are joining Twitter to share their stories, express their opinion about different topics, create content marketing etc.

2.1.1. Library for Collecting Tweets API

To collect the tweets, `searchtweets`, a Python library created by Twitter that serves as a wrapper for search APIs of the enterprise and premium licenses to provide a command-line utility, was employed [12]. This library provides the following useful features:

- Outputs a stream of data for low in-memory requirements
- 30-day Search (including Full Archive Search)
- Flexibility of use in Python programs
- Premium and Enterprise authentication methods

The news tweets are collected from the Malaysia News Twitter accounts, namely The Star, New Straits Times, The Edge Malaysia, The Sun daily, Malay Mail, BFM Radio, TheEdge Property, and The Star R.age, using Twitter API.

2.2. Programming Language Chosen

Python is being chosen to conduct this research. Python is a high level, general-purpose programming language that is user-friendly in which it emphasizes the readability where most of the keywords are plain English [13]. It is an interpreted programming language where it can be executed directly without needing to be compiled down into machine-language instructions before executed by the machine.
2.3. Text Preprocessing Library

NLTK is a Python natural language processing (NLP) library to handle human language data [14]. It has easy-to-use interfaces with over 50 corpora and lexical resources, and a suite of text processing libraries for tokenization, tagging, classification, stemming, semantic reasoning, parsing, supported with an active discussion forum [15].

SpaCy is a free, open-source library for advanced NLP in Python [16]. SpaCy is built on Cython, a superset of Python that has C-like performance, hence it is fast. It is designed specifically for production use and helps to build an application that processes and “understand” large volumes of text. Both NLTK and SpaCy were used to clean the text by removing stop words, punctuations and stemming of words to remove the noise in the data. Example of stop words include are, an, the, while, etc. that only serve as conjunction in the sentence.

2.4. Library for Machine Learning

Python has many matured machine learning libraries to perform the classification process. The libraries used were TensorFlow, Keras, Gensim and Scikit-learn.

Tensorflow provides a comprehensive flexible ecosystem of libraries, tools and community resources, to build and deploy machine learning-powered applications. It was developed by the Google Brain Team in 2015 [17]. TensorFlow provides an extensive machine learning library to develop and train models in Python, JavaScript, C++, Java, Go language and Swift. TensorFlow also has simplified the flow to deploy the model in the cloud, on-premise, in the browser, or on-device through the collection of tools in the library [18]. The model that builds and trains in TensorFlow is run as a blackbox as illustrated in Fig. 2 [19]. However, the granularity that TensorFlow offers has become a barrier for newcomers.

![Fig. 2: Black box model in TensorFlow](image)

Keras is TensorFlow's high-level API for building and training deep learning models, used for state-of-the-art research, fast prototyping and production [20]. The advantages of Keras over TensorFlow are:

- User-friendly as it has a simple and consistent interface that is optimized for common use cases. It also has clear and actionable feedback on user errors.
- Modular and composable by connecting configurable building blocks together to build he models, with few restrictions.
- Easy to extend as it allows custom building blocks to be written to express new ideas for research. Examples include the development of new metrics, layers, loss functions, and state-of-the-art models.

The machine learning model that is trained in either TensorFlow or Keras can be deployed in the web browser by using TensorFlow.js.
Gensim supports document indexing, topic modeling, and similarity retrieval with large corpora [21]. It is another Python library that is well-optimised for NLP, information retrieval (IR) and document similarity analysis [22].

Scikit-learn for machine learning, is a Python module built on top of SciPy [23]. Designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy, it provides support for regression, classification and clustering algorithms including random forests, support vector machines (SVM), k-means, gradient boosting and DBSCAN.

3. Results and Discussion

3.1. Supervised Learning

The accuracy of the models is the performance evaluation of the performance of the models. The higher the better. ANN obtained the lowest accuracy score among the models in classifying the tweets, while the highest accuracy was achieved by the bidirectional Stacked LSTM with the accuracy score of 92.54%. The second lowest accuracy was logistic regression with a score of 30.28%, followed by SVM, RNN, LSTM, Stacked LSTM, and Bidirectional LSTM with the accuracy score of 33.64%, 48.93%, 57.19%, 60.24% and 60.55%, respectively. This performance observation was due to the ambiguity of the words in the sentence that had increased the difficulty in classifying the tweets. Moreover, as tweets are of very limited length, it requires “common sense” to understand the concept of the tweets within a few words. ANN, logistic regression and SVM do not have the embedding layer to help the model to extract the hidden pattern from the tweets. Whereas, RNN and LSTM have the additional embedding layer to store the corpus and extract the hidden meaning throughout the training phase.

Logistic regression and SVM were able to outperform ANN by a margin as ANN uses the weight of each neuron to classify the tweets, while logistic regression and SVM create a separating line to classify the tweets. The hyperplane in SVM and polynomial function in logistic regression are less affected by the ambiguity of the words. SVM has higher accuracy than logistic regression because logistic regression tries to find a best fit line while SVM finds the best space to form a hyperplane to separate the classes. As there are more than 3 thousand points, fitting the point with the best fit line is harder than drawing a line in the space to separate the classes.

Stacked bidirectional LSTM was able to score the highest accuracy among other models because the accuracy of the neural network often increases as the depth of the model increases. This is why the accuracy of the stacked LSTM is similar to bidirectional LSTM. However, when the depth of the bidirectional LSTM increases, it can lead to a significant improvement in the accuracy. In comparison, the ordinary LSTM only see inputs from the past, while bidirectional LSTM learn from two way inputs: past to future, and future to past [24]. Moreover, with the addition of layers, the hidden neuron is able to add levels of abstraction of input observations over time; thus the model is able to chunk observations over time or represent the problem at different time scales to achieve a better accuracy [25].

3.2 Unsupervised Learning Model

The metrics to evaluate the LDA is text perplexity. For the purposes of the evaluation, the lower the perplexity, the better.

The Gensim LDA model was evaluated via coherence score and text perplexity. The Gensim LDA model achieved low text perplexity of -23.211504971215284, which is the probability distribution of the sentences or texts in the model. Hence, theoretically this is a good model. The visualisation of the word per topic, as shown in Fig. 3, indicates that the topics overlap and hence the low perplexity score of the model.

Gensim LDA also supports topic modeling using bag of words (BOW) only. Gensim BOW LDA obtained a text perplexity score of -9.240800339985618 and has lower overlapping of the topic as shown in Fig. 4. This shows that the presence of an e-dictionary will cause the Gensim LDA model to over-map the word to the dictionary of words, thus causing the overlapping of the topic in the model.
(a) Inter-topic Distance Map

(b) Top 30 Words of the topic
The text perplexity of the Scikit-learn LDA is higher than Gensim LDA with the score of 2268.962449476619. However, Fig. 5 shows that there is a distinct distribution of the topic. Therefore,
the text perplexity cannot be the only indicator to evaluate the performance of the model. Rather, there is also a need to factor in the distance between the topics.

4. Conclusion

The objectives of the work had been achieved. Stacked bidirectional LSTM is able to achieve 92.54%, which was considered a high-performance model. For unsupervised model to look for the hot keywords will be scikit-learn LDA model as it can classify the keywords into the topic without overlapping with each other.

References

[1] K. Leetaru, "Is Twitter Really Faster Than The News?", Forbes, 2019. [Online]. Available: https://www.forbes.com/sites/kalevleetaru/2019/02/26/is-twitter-really-faster-than-the-news/#6cb033381ef7. [Accessed: 04- Apr- 2020]

[2] "Twitter: number of users in Malaysia 2019 | Statista", Statista, 2020. [Online]. Available: https://www.statista.com/statistics/490591/twitter-users-malaysia/. [Accessed: 04- Dec- 2020]

[3] "Report: More Malaysians going online for news fix", Free Malaysia Today, 2017. [Online]. Available: https://www.freemalaysiatoday.com/category/nation/2017/06/22/report-more-malaysians-going-online-for-news-fix/. [Accessed: 04- Apr- 2020]

[4] Brownlee, "Logistic Regression for Machine Learning", Machine Learning Mastery, 2019. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. [Accessed: 29-Nov-2019]

[5] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3, 1999.

[6] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms", Medium, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed: 05- Apr- 2020]

[7] A. Krenker, J. Bester and A. Kos, "Introduction to the Artificial Neural Networks", Artificial Neural Networks - Methodological Advances and Biomedical Applications, 2011. [Accessed: 27-Nov-2019]

[8] A. Shekhar, "Understanding The Recurrent Neural Network", Medium, 2018. [Online]. Available: https://medium.com/mindorks/understanding-the-recurrent-neural-network-44d593f112a2. [Accessed: 18- Nov- 2019]

[9] "Understanding LSTM Networks -- colah's blog", Colah.github.io, 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 19- Dec- 2019]

[10] Blei, D.M., Ng, A.Y., and Jordan, M.I. “Latent dirichlet allocation. Journal of machine learning research”, pp.993–1022, 3-Jan-2003. [Online] Available: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [Accessed: 19- Dec- 2019]

[11] A. Java, X. Song, T. Finin and B. Tseng, "Why We Twitter: An Analysis of a Microblogging Community", Advances in Web Mining and Web Usage Analysis, pp. 118-138, 2009.

[12] twitterdev/search-tweets-python, GitHub, 2018. [Online]. Available: https://github.com/twitterdev/search-tweets-python. [Accessed: 28- Dec- 2019]

[13] "Compiled Versus Interpreted Languages", Guide.freecodecamp.org, 2020. [Online]. Available: https://guide.freecodecamp.org/computer-science/compiled-versus-interpreted-languages/. [Accessed: 27- Dec- 2019]

[14] Nltk.org. 2019. Natural Language Toolkit — NLTK 3.5B1 Documentation. [online] Available at: <https://www.nltk.org/> [Accessed 8 January 2020].
[15] Bird, S., Klein, E. and Loper, E., 2011. Natural Language Processing With Python. Beijing [etc.]: O'Reilly.

[16] "Industrial-Strength Natural Language Processing in Python", Spacy.io, 2020. [Online]. Available: https://spacy.io/. [Accessed: 03- Jan- 2020]

[17] Google Inc., "The Google Brain Team — Looking Back on 2017 (Part 1 of 2)", Google AI Blog, 2020. [Online]. Available: https://ai.googleblog.com/2018/01/the-google-brain-team-looking-back-on.html. [Accessed: 04- Jan- 2020]

[18] Google Developers, "API Documentation | TensorFlow Core v2.1.0", TensorFlow, 2020. [Online]. Available: https://www.tensorflow.org/api_docs. [Accessed: 03- Jan- 2020]

[19] A. Kumar, "QA: Blackbox Testing for Machine Learning Models - DZone AI", dzone.com, 2018. [Online]. Available: https://dzone.com/articles/qa-blackbox-testing-for-machine-learning-models. [Accessed: 09- Feb- 2020]

[20] S. Gupta, J. Gordon and K. Allison, "Standardizing on Keras: Guidance on High-level APIs in TensorFlow 2.0", Blog.tensorflow.org, 2018. [Online]. Available: https://blog.tensorflow.org/2018/12/standardizing-on-keras-guidance.html. [Accessed: 13- Jan- 2020]

[21] Řehůřek, R. and Sojka, P., “Software Framework for Topic Modelling with Large Corpora. English.” Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, pp.45–50, 2010.

[22] scikit-learn, "scikit-learn/scikit-learn", GitHub, 2019. [Online]. Available: https://github.com/scikit-learn/scikit-learn. [Accessed: 25- Nov- 2019]

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel; P. Prettenhofer; R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, 2011.

[24] Stack overflow, "What's the difference between a bidirectional LSTM and an LSTM?", Stack Overflow, 2017. [Online]. Available: https://stackoverflow.com/questions/43035827/whats-the-difference-between-a-bidirectional-lstm-and-an-lstm. [Accessed: 02- Mar- 2020]

[25] J. Brownlee, "Stacked Long Short-Term Memory Networks", Machine Learning Mastery, 2019. [Online]. Available: https://machinelearningmastery.com/stacked-long-short-term-memory-networks/. [Accessed: 02- Mar- 2020]