淡江大學資訊管理學系碩士班

碩士論文

指導教授：鄭啟斌 博士

應用深度學習與自然語言處理於仇恨言論之自動偵測

Apply Deep Learning and Natural Language Processing to Hate Speech Automatic Detection

研究生：蔡坤利

中華民國 109 年 1 月

論文名稱：應用深度學習與自然語言處理於仇恨言論　　頁數：64

校系(所)組別：淡江大學資訊管理學系碩士班

畢業時間及提要別：108 學年度第 1 學期碩士學位論文提要

研究生：蔡坤利　　指導教授：鄭啟斌 博士

論文提要內容：

隨著網路的發展，社群媒體使用人數也逐年攀升，網路仇恨言論的問題也伴隨著發生，這個問題的影響不僅僅存在於網路，甚至影響網路使用者的身心狀況。僅管社群媒體管理方已投入大量人力與金錢試圖解決這個問題，然而仍被使用者認為成效不彰。而本研究透過深度學習與自然語言處理，使用了兩個不同的資料集，皆為內含標記仇恨言論的 Twitter 推文，使用了兩個深度學習模型：BERT 模型與 Bi-LSTM 模型，透過深度學習的方式去預測 Twitter 推文是否為仇恨言論。本研究結果顯示，使用 BERT 模型進行仇恨言論偵測的成效較優於使用 Bi-LSTM 模型，本研究也發現，資料集內仇恨言論所佔的比例，將會影響到使用深度學習模型預測的結果。

關鍵字：BERT;自然語言處理;仇恨言論;深度學習

Title of Thesis:	Apply deep learning and natural language processing to hate speech automatic detection	Total pages:64
Key word:	BERT, NLP, hate speech, deep learning	
Name of Institute:	MASTER'S PROGRAM DEPARTMENT OF INFORMATION MANAGEMENT, TAMKANG UNIVERSITY	
Graduate date:	January, 2020	Degree conferred: Master
Name of student:	Kun-Li Tsai 蔡坤利	Advisor: Dr. Chi-bin Cheng 鄭啟斌 博士

Abstract:

With the development of the Internet, the number of social media users has also increased year by year, and the problem of "hate speech" on the Internet has also occurred. The impact of this problem not only exists on the Internet, but also affects the physical and mental conditions of Internet users. Although social media companies have invested a lot of manpower and money in trying to solve this problem, they are still considered ineffective by users. This study uses deep learning and natural language processing to use two different data sets, both of which are Twitter tweets containing labeled hate speech. Two deep learning models are used: the BERT model and the Bi-LSTM model. Learn ways to predict whether the Twitter tweets are hate speech. The results of this study show that the performance of hate speech detection using the BERT model is better than that of the Bi-LSTM model. This study also found that the proportion of hate speech in the data set will affect the prediction results using the deep learning model.

目錄

圖目次

表目次

第一章 緒論

隨著網路的發展，加上各式行動上網裝置的普及，社群網站的使用人數也逐年攀升（如表 1 所示），由社群媒體數字營銷機構 DataReportal 於 2019 年所提出之 Digital 2019: Global Digital Overview[1]報告指出，在全球 76.76 億人口中，有 34.83 億人使用社群媒體，佔了全球人口的 45%，不僅是一般社群網站，線上影片串流網站、直播平台、新聞網站、部落格文章也都提供留言或即時聊天室的功能，藉此，人們能以更快速、即時且方便的方式發表言論或獲得訊息，也因此，人們和網路社群的關係也變得更加緊密。

表 1 全球社群媒體使用者人數 （Digital 2019: Global Digital Overview, 2019）

	2014	2015	2016	2017	2018	2019
人數（億）	18.57	20.78	23.07	27.96	31.96	34.84
與前一年變化		+12%	+11%	+21%	+14%	+9%

儘管網路社群帶給人們這些好處，問題也伴隨著發生，由於網路社群發展迅速與其匿名性與任何人皆可發言的特性，使得仇恨言論（hate speech）的散播也更加容易。仇恨言論是指特定形式會引起他人反感的言論，通常被定義為針對某些特性如種族、膚色、文化、性別、性向、國籍、宗教等，來貶低個人或團體的言論(Nockleby, 2000)。由美國獨立智庫機構 Pew Research Center 於 2018 年所提出的研究指出，在美國有 59%的青少年曾在網路上經歷某些形式的霸凌（如表 2

[1] Kemp, S. （2019）. DIGITAL 2019: GLOBAL DIGITAL OVERVIEW. Retrieved from https://datareportal.com/ reports/ digital-2019-global-digital-overview

所示），受訪青少年中有 90%的人認為這是個會影響他們這年紀的人的問題，然而，大多數受訪青少年認為他們的老師、社群媒體網站管理方及政府單位並不重視這類問題，這類仇恨言論將導致人們在社群媒體上接受到不正確的訊息以及觀念，對人們自身心態，甚至整個社會造成負面的影響。

表 2 美國青少年在網路經歷網路霸凌比例

（A Majority of Teens Have Experienced Some Form of Cyberbullying , 2018）[2]

Any type of cyberbullying listed below	59%
Offensive name-calling	42%
Spreading of false rumors	32%
Receiving explicit images they didn't ask for	25%
Constant asking of where they are, what they're doing, who they're with, by someone other than a parent	21%
Physical threats	16%
Having explicit images of them shared without their consent	7%

過去幾年，為避免上述仇恨言論造成的問題，社群網站如 Facebook, YouTube, Twitter 等也投入大量資金與人力試圖去解決(Lomas, 2017)，但仍被使用者認為效果不佳(Gambäck & Sikdar, 2017)，因為這些網站過去大多透過其他用戶的手動檢舉、人工審核或關鍵字的屏蔽來處理網路留言，這類方法需倚靠大量人力且耗時，已無法面對現今快速且大量的網路留言(Waseem & Hovy, 2016)。然而，這些社群網站目前也開始使用機器學習的方式來偵測仇恨言論，不過成效仍

[2] Anderson, M. （2018）. A Majority of Teens Have Experienced Some Form of Cyberbullying. Retrieved October 28, 2019, from https://www.pewinternet.org/ 2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

因訓練數據的偏差，而導致這些深度學習模型的準確率降低(Davidson, Bhattacharya, & Weber, 2019)。2017 年，德國政府也為此批准法案，要求社群媒體必須對明顯違法之言論或有爭議之言論在期限之內做處理，並將結果告知投訴者，若沒有做到將對社群媒體罰款。

為此，本研究利用自然語言處理（Natural Language Processing, NLP）和深度學習（Deep Learning）結合，使用兩個仇恨言論資料集作為訓練語料，皆包含社群網站 Twitter 推文，與是否為仇恨言論之標記，使用兩個深度學習模型，BERT 模型與雙向長短期記憶（Bi-LSTM）模型，讓模型透過深度學習，判斷該自然語言是否為仇恨言論，並將兩個模型的結果進行比較與分析。

儘管仇恨言論對人們的影響甚巨，但在數量龐大的網路留言中還是佔極小的部分(Zhang & Luo, 2018)，在這種資料集極度不平衡的情況之下，即使偵測出的少數群體（仇恨言論）數量有明顯上升或下降，若我們只關注於模型整體正確率可能無法看出正確成效，因此，本研究也將偵測結果的精準率（precision）、召回率（recall）和 F1-score 列為評估標準。期望藉由本研究，分析與說明使用這兩個不同的深度學習模型於相同的仇恨言論資料集偵測結果有何差異、提高仇恨言論自動偵測準確率，以利社群媒體對於仇恨言論的管理，降低對社會的負面影響。

本研究之研究流程如下（圖 1）:

1. 研究動機與目的

   確立本研究題目目前之限制，與可能可改進之方向。

2. 相關文獻探討

   參考過去將自然語言處理（NLP）與深度學習結合之相關研究，藉此確立深度學習模型之使用與可改進之方向。

3. 模型建立與修改

確定本研究欲使用之兩個模型後，使用 Python3 語言建置，並做參數與模型架構之調整，以找出適合之組合。

4. 評估與結果分析

運用正確率、精準率、召回率和 F1-score 作為評估標準，與過去相關研究做比較與分析。

5. 結論與建議

針對本研究做出結論，提出不足與可改進之處。

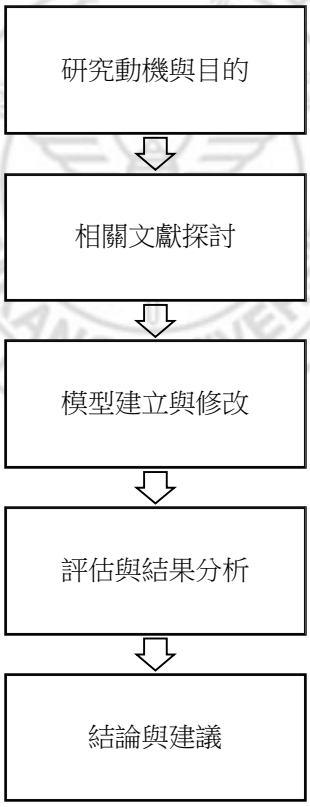


圖 1 研究流程

第二章 文獻探討

在本章節，我們將探討過去與仇恨言論偵測相關的研究，探討他們使用的方法、與使用的術語，並從中瞭解差異。

2.1 仇恨言論

仇恨言論，是一個可以概括大多數污辱性言論的詞彙，也是目前相關研究中最常使用的詞，甚至是一些國家的法律用詞(Schmidt & Wiegand, 2017)。而在過去的研究中，也有人使用相似的詞彙表示，(Spertus, 1997)使用了辱罵訊息（abusive messages）、敵意訊息（hostile messages）和謾罵（flames），而近期也有許多研究使用網路霸凌（cyberbullying）這個詞(Dadvar, Trieschnigg, Ordelman, & de Jong, 2013; Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Hosseinmardi et al., 2015; Van Hee et al., 2015; Xu, Jun, Zhu, & Bellmore, 2012; Zhong et al., 2016)。仇恨言論一詞，則是由(Warner & Hirschberg, 2012)開始使用。也有研究(Sood, Antin, & Churchill, 2012)使用污辱（insults）、褻瀆（profanity）和惡意（malicious intent）和(Razavi, Inkpen, Uritsky, & Matwin, 2010)使用的攻擊性言語（offensive language）。

判斷一段留言或文章是否為仇恨言論並不容易，即便交給人工判斷也有可能因為不同的文化、語言或情境而有不同的判斷(Fortuna & Nunes, 2018)，因此，仇恨言論的定義變得很重要，我們整理出其他研究對於上述與仇恨言論相關詞彙的定義（如表 3），也整理出了歐盟（European Union, EU）和社群媒體 Facebook、Twitter、YouTube 對於仇恨言論的定義（如表 4）。

表 3 其他研究對於相似詞彙的定義

詞彙	定義
網路霸凌（Cyberbullying）	透過電子形式反覆的對無法輕易為自己辯護的個人或團體進行的故意傷害行(Chen, 2011)。
歧視（Discrimination）	不公平的對待別人的差異(Thompson, 2016)。
謾罵的（Flaming）	可能會影響社群的煽動性且充滿敵意、謾罵、恐嚇性的言論(Guermazi, Hammami, & Hamadou, 2007)。
辱罵性言語（Abusive language）	指傷害性言語，包括仇恨言論、謾罵性言論和貶低性言論(Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016)。
褻瀆（Profanity）	令人反感的或淫穢的詞彙或片語[3]（Vigna et al., 2017）。
令人不快的言語或評論（Toxic language or comment）	不禮貌、不尊重或不合理的訊息，可能會使其他使用者放棄討論[4]（Jigsaw, 2017）。
極端份子、觀點（Extremism）	與極端份子或仇恨團體相關的意識形態，助長暴力活動，通常目的為撕裂族群、重新取得社會地位，這群人通常有犯罪前科或處於社會底層。(McNamee, Peterson, & Peña, 2010)。
激進（Radicalization）	與極端主義類似，已有許多研究是針對這類領域如恐怖主義、反黑人社群和民族主義(Agarwal & Sureka, 2015)。

[3] Cambridge Dictionary. 2017. Profanity. Retrieved from https://dictionary.cambridge.org/dictionary/english/profanity.

[4] Jigsaw. 2017. Perspective API. Retrieved from https://www.perspectiveapi.com/.

表 4 歐盟和社群媒體對仇恨言論的定義

社群媒體	定義
歐盟[5]	針對某些特性如種族、膚色、宗教、血統、國籍、文化等，來貶低個人或團體的言論
Facebook[6]	針對他人的種族、民族、國籍、宗教、性傾向、社會地位、性別、性別認同、重大疾病或身心障礙等所謂的保障特徵進行直接攻訐。我們也為移民身分提供一些保護措施。我們對攻擊的定義，包括暴力或非人化的言論、貶抑的陳述方式，或鼓吹排擠或隔離。
Twitter[7]	您不得基於種族、民族、國籍、性傾向、性別、性別認同、宗教派別、年齡、殘疾或疾病，助長對其他人的暴力，或是直接攻擊或威脅其他人。我們也不允許任何帳戶基於上述類別來煽動對他人的傷害行為。
YouTube[8]	基於下列任一項特質，鼓吹對個人或群體採取暴力行為或煽動仇恨，我們就會將其移除：年齡、種姓制度階級、身心障礙、族裔、性別認同、國籍、種族、移民身分、宗教、性別/性別氣質、性傾向、重大暴力事件的受害者及其親屬、退役身分。

(Fortuna & Nunes, 2018)的研究整理了 128 篇與仇恨言論相關文獻，並為其分類，其中最多的一類為「法律與社會科學」，總計有 76 篇文獻，顯示在法律領域、社會科學領域對於仇恨言論的研究數量是多於電腦科學領域的。而與仇恨言論偵測相似的網路霸凌，在社會科學與心理學領域也有許多相關的研究，許多社會科學對此方面的研究都致力於評估網路霸凌的流行程度，特別是針對兒童跟青少年(Mishna, Cook, Saini, Wu, & MacFadden, 2011)。精神病學的研究則探討網路

[5] Wigand, C., Voin. M., （2017）. Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law. Retrieved from http://europa.eu/rapid/ press-release_SPEECH-17-403_en.htm.

[6] Facebook. 2013. What does Facebook consider to be hate speech? Retrieved from https://www.facebook.com/help/135402139904490.

[7] Twitter. 2017. The Twitter Rules. Retrieved from https://support.twitter.com/articles/.

[8] Youtube. 2017. Hate speech. Retrieved from https://support.google.com/youtube/answer/2801939?hl=en.

霸凌對於兒童及青少年長期及短期的影響，以及父母、教育者和精神健康工作者的處理方式(Patchin & Hinduja, 2012)。(Dinakar et al., 2012)的研究則提到，上述這類研究通常涉及密集的調查與訪談，並且提供了一些重要的演算法見解，對於改進網路霸凌偵測的模型有幫助。

2.2 仇恨言論偵測使用的特徵

過去的研究做仇恨言論偵測時，使用許多不同特徵作分類工作，因為將仇恨言論與非仇恨言論區分的方式往往不是受單一的因素所影響，使用不同的特徵作分類工作，會有不同的成效與結果，本節將探討過去相關的研究使用的不同的分類特徵。

2.2.1 表面特徵（Surface Features）

對於文字分類來說，最容易拿來使用的就是文字的表面特徵，像是使用詞袋（bag of words）或是單詞或字母的 n-gram，有許多研究使用這些簡單的表面特徵作仇恨言論偵測(Pete Burnap & Williams, 2015; Davidson, Warmsley, Macy, & Weber, 2017; Del Vigna, Cimino, Dell'Orletta, Petrocchi, & Tesconi, 2017; Kwok & Wang, 2013; Waseem & Hovy, 2016)，攻擊性、辱罵性言論(Chen, Zhou, Zhu, & Xu, 2012; Mehdad & Tetreault, 2016; Nobata et al., 2016)、歧視性(Yuan, Wu, & Xiang, 2016)和網路霸凌(Zhong et al., 2016)的偵測，而使用這類特徵通常也能得到不錯的成效，不過(Nobata et al., 2016)也提到，雖然單詞或字母的 n-gram 是他們研究中使用單一特徵預測中成效最好的，但若將他們與其他特徵結合使用，會得到更好的預測成效。

而網路上的留言中，常常會遇到有些詞不按照正規拼法，而使用特殊的符號或數字代表字母，例如這句話“ki11 yrslef a$$hole”，將字母 S 改成符號$或是將字母 l 改成數字 1，還有拼字時漏掉不影響閱讀的字母，這類特殊拼字對其他使用者理解句子可能不會造成困擾，但若是在文字分類上使用單詞的 n-gram 特徵時便會造成問題，而使用針對字母的 n-gram 特徵可能是這問題的解決方式(Mehdad & Tetreault, 2016)，他們的研究比較了單詞的 n-gram 特徵和字母的 n-gram 特徵用於仇恨言論偵測，結果顯示字母的 n-gram 預測效果較佳。除了使用單詞或字母的 n-gram 特徵外，其他的文字表面特徵如標點符號、字母大小寫、詞典中找不到的詞彙等特徵，也對仇恨言論偵測的成效有些許的提升(Chen et al., 2012; Nobata et al., 2016)。

2.2.2 詞彙一般化（word generalization）

雖然使用詞袋於仇恨言論偵測有不錯的成效，不過這類方式需要特徵同時出現在測試集與訓練集，而仇恨言論通常為一段短文甚至單一句子，因此有可能會出現數據稀疏的問題(Schmidt & Wiegand, 2017)，為解決這問題，有些研究開始使用聚類，將詞彙分到不同的類別，並以各類別的標籤作為特徵，這種將詞彙一般化的方式，布朗聚類（Brown clustering）就是其中一個標準的演算法，這個演算法最初會將每一個詞分成獨立的一類，之後再將兩個類別合併，重複合併直到設定的類別數量時停止，(Warner & Hirschberg, 2012)的研究就是以此聚類方式做為特徵。

而近期的研究中也有人開始使用詞嵌入（word embedding 或 word representation）的方式，此方式為將詞轉換為詞向量（word vector），以詞向量作

為特徵，也就是在一個文本中，針對每個單詞，給出一個向量來表示他(Mikolov, Chen, Corrado, & Dean, 2013)，這樣一來，便能將一句話，分成一個個單詞的詞向量表示，以此作為神經網路的輸入。詞嵌入在過去的作法大多為使用 one hot encoding，而此做法類似將文本中出現的字一一列入字典做編號，而這個向量的維度就是該文本中詞的數量，而那代表一個詞的向量中，只有一個維度的值為 1，其他所有的值都是 0，假設該文本中有 n 個不重複的詞，而代表每一個單詞的向量中就會有 1 個 1 和 n-1 個 0。

one hot encoding 雖然直觀且容易使用，但卻沒辦法表示出詞與詞之間的關聯性(李洋 & 董红斌, 2018)。為此，(Mikolov et al., 2013)基於(Bengio, Ducharme, Vincent, & Jauvin, 2003)所提出的 Neural Network Language Model（NNLM）提出了 word2vec 模型。Facebook 在 2015 年也提出 fastText，還有(Pennington, Socher, & Manning, 2014)提出的 GloVe 和(Ji, Yun, Yanardag, Matsushima, & Vishwanathan, 2015)提出的 WordRank，其中 word2vec 模型還是最多研究使用的。

word2vec 模型將單詞轉換為高（20-300）維的向量，並提出了 Continuous Bag-of-Words（CBOW）和 Skip-gram 兩種訓練詞向量的方式，而兩種訓練方式的差異在於，Skip-gram 則是透過中間的詞來預測上下文，假設我們有一個輸入詞W_t，將 skip_window 參數設為 2，表示模型將會選取輸入詞的前 2 個詞W_{t-1}, W_{t-2}及後兩個詞W_{t+1}, W_{t+2}，包含輸入詞模型總共得到五個詞W_{t-1}, W_{t-2}, W_t, W_{t+1}, W_{t+2}，若 skip_window 參數大於輸入詞前、後詞的數量，只會取到句首或句尾，而 CBOW 則是完全相反，透過上下文來預測中間的詞，(如圖 2)，透過 word2vec 模型訓練後，只要是詞義相近，即使是不同的詞，也會得到相近的詞向量，如此一來，我們便可以透過詞與詞之間向量的距離，來表示詞的關聯性。

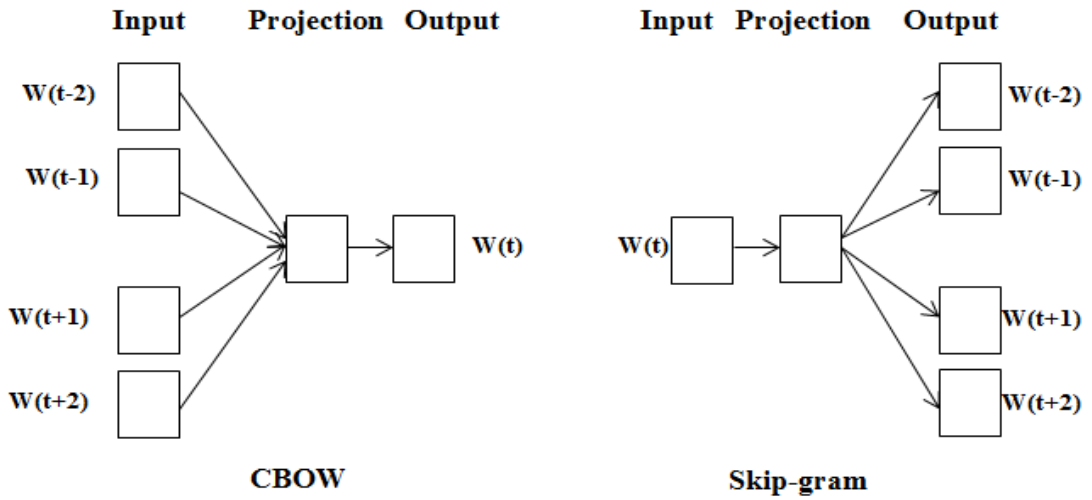


圖 2 Continuous Bag-of-Words（CBOW）,Skip-gram(Mikolov et al., 2013)

由於仇恨言論通常為一個短文或一句話，在獲得詞向量以後，必須要得到該短文或那句話的單詞詞向量集合來表示，最簡單的方式便是將該短文或句子內的單詞詞向量求平均，然而這方法不論是使用普通的方法或是使用特定領域的語料庫對單詞做預訓練，效果都十分有限(Nobata et al., 2016)。

2.2.3 情感分析（Sentiment Analysis）

因為負面情緒（negative sentiment）言論和仇恨言論很類似，所以情感分析和仇恨言論偵測的研究也是高度相關的，因此，過去有許多關於仇恨言論偵測的研究是以不同情感為特徵去做分類研究(Schmidt & Wiegand, 2017)，(Van Hee et al., 2015)就曾以情感詞彙，對文本中的詞彙進行分類，分成正面（positive）、負面（negative）及中性（neutral）詞彙，以此為特徵做監督式學習分類，而這類研究屬於單步驟方法，也有多步驟方法，(Dinakar et al., 2012)、 (Sood et al., 2012)和 (Gitari, Zuping, Damien, & Long, 2015)的研究在判斷是否為仇恨言論的分類器

前，須先經過判斷是否為負極性（negative polarity）詞的分類器，而(Gitari et al., 2015)則在前述研究中判斷負極性詞的分類器前，再加上一個主觀、非主觀（non-subjective）的分類器。

2.2.4 詞彙資源（Lexical Resources）

仇恨言論通常被認為會包括一些特定的批評或是污辱的否定詞，要獲得這類否定詞語料，必須要找到相關的詞彙資源，而網路上就有一些公開的詞彙資源，裡面包含了與仇恨言論相關的術語，許多研究(Pete Burnap & Williams, 2015; Nobata et al., 2016; Xiang, Fan, Wang, Hong, & Rose, 2012)都依此類詞彙作為特徵，也有一些詞彙資源是專門提供針對特定領域、族群像是種族歧視、性別歧視、LGBT 用語、殘障人士的否定詞。除了網路上公開的詞彙資源，有些研究也自行編輯了仇恨言論的詞彙資源，(Razavi et al., 2010)自行編輯了《侮辱性語言詞典》，該辭典收集了 2700 個侮辱性詞彙與片語，並依照對仇恨言論偵測研究的淺在影響程度，透過訓練給予各詞彙或片語一個權重。大多數的研究都將詞彙資源作為輔助特徵或是其他用途，與其他特徵像是 bag of words 或是詞嵌入相比，詞彙資源通常不足以單獨作為仇恨言論偵測的特徵使用(Nobata et al., 2016)，這是因為即使使用了這些污辱性詞彙或片語，我們還是必須透過瞭解他使用的情境與背景才能判斷是否為仇恨言論，(Hosseinmardi et al., 2015)的研究便指出，儘管新聞媒體報導使用侮辱性詞彙或片語的比例相當高，但在這之中仍有 48%不被認為是仇恨言論。

2.2.5 語言學特徵（Linguistic Features）

語言學對於仇恨言論偵測也有些幫助，因此，許多研究也將語言學特徵加入使用，(Xu et al., 2012)的研究將 n-gram 特徵與 POS（part-of-speech）資訊結合，POS 資訊也就是詞性、詞類、句法類別的資訊，有了這類資訊可以讓我們對句子有更多的了解，例如知道一個單詞的詞性為動詞或名詞，我們就可以推定在他前後相鄰的單詞可能的詞性，像是動詞前面可能為名詞，儘管如此，在他們的研究結果顯示，加入 POS 資訊對仇恨言論偵測的幫助不大。

而(Chen et al., 2012)的研究著重於更深層的句法訊息，特別是針對較長的文本，使用類型的依存關係，這個依存關係可以在同一特徵中找出不連續的但兩者有關係的詞組，像是豬（pigs）跟猶太人（Jews），兩個詞若是不使用依存關係且在句子中不為連續詞，將無法看出這兩個詞彙的關係，使用依存語法表示這兩個詞彙：nsubj（pigs, Jews），便可看出 pigs 和 Jews 有依存關係，類型為 nsubj（nominal subject），即是名詞主語關係，依此表示冒犯性言語 pigs 與攻擊目標 Jews 之間的關係，其他還有 dobj（direct object），直接受詞關係、iobj （indirect object），非直接受詞關係等不同的依存關係，以此方式，進而找出更多隱性的仇恨言論。

而(Zhong et al., 2016)的研究並沒有使用句子中詞彙的依存關係作為特徵，而是使用冒犯性等級分數（offensiveness level score）作為特徵，這個分數是建立於相同依存關係中同時出現冒犯性詞彙與用戶標識符的頻率。

(Spertus, 1997)在他的 Smokey 系統研究中，使用語言學特徵設計了一個仇恨言論偵測的系統，其中包含了對於強制性語句的偵測。而 Smokey 系統還結合了一些句法特徵來防止錯誤預測（False Positive），像是建立一些規則如：讚美規則

（praise rules），如果該句子不包含污辱性詞彙，則歸類入讚美類，另外一個簡單的讚美類規則則是，句子若中出現他們預先定義的好詞（good words），像是「祝福」、「恭喜」、「榮譽」、「一路順風」之類的詞彙，則歸類至讚美類，其他還有禮貌規則（Polite Rules）、污辱（Insults）、謙虛規則（Condescension Rules）、褻瀆規則（Profanity Rules）等各式不同規則。(Nobata et al., 2016)的研究也使用了類似的特徵。

2.2.6 知識庫（Knowledge-Based）特徵

有些言論，我們無法直接的從字面使用了哪些關鍵字上來判斷是否為仇恨言論，(Dinakar et al., 2012)的研究便指出，若要解決這種隱性的仇恨言論，必須要將常用的刻板印象和社會觀念建立一個知識庫，該研究從 Formspring 語料庫中舉出一句話做為為範例。

「戴上假髮和畫上口紅，做你真正的自己。」

這個例句中，我們並沒有辦法直接看出他是否為仇恨言論，會因為這句話告知的對象不同而有不同的判斷，假設這句話告知對象為一名異性戀男性，因假髮和口紅為對女性常見的刻板印象，而異性戀男性可能會因傳統社會觀念，不喜歡被賦予相反性別的特徵，不過，要是這句話為同性戀者之間的對話，就有可能是無傷害性的。(Dinakar et al., 2012)便使用 Formspring 資料集，建立了一個針對性別、LGBT 刻板印象的知識庫（如圖 3）。

只有女生會使用口紅
口紅是化妝的一部分
只有女生會化妝
只有女生會使用假髮
遮頭頂圓禿的假髮只有男生會使用

圖 3 性別、LGBT 刻板印象的知識庫範例

該知識庫收集了超過 200 條斷言，以與 ConceptNet 相同的方式將他們轉換為概念與關係的稀疏矩陣表示，ConceptNet 為美國麻省理工（MIT）提出的語意網路，其中包含大量這世界的資訊，以供電腦使用。不過該研究建立的知識庫還是只對於特定領域（性別、LGBT）有幫助。

2.2.7 元資料（Meta-Information）

除了知識庫，元資料（關於此言論的資料）也是仇恨言論偵測可以使用的資料，由於大部分的仇恨言論偵測使用的資料集都是來自於社群媒體，透過社群媒體提供的 API 也可額外獲得關於留言的其他資訊，像是發言者的性別、年齡、地理位置之類不同的背景資料，獲得發言者的背景資料對於仇恨言論的偵測是很有幫助的，(Schmidt & Wiegand, 2017)。(Xiang et al., 2012)的研究就是使用上述推論偵測到更多的仇恨言論。(Dadvar et al., 2013)的研究則是將使用者發文中污辱性詞彙的數量做紀錄，以此作為特徵。(Dadvar et al., 2013; Waseem & Hovy, 2016)的研究則表示，男性發表仇恨言論的可能性較女性高，由此可知，知道發言者的性別對仇恨言論的偵測也是有幫助的。

(Schmidt & Wiegand, 2017)也提到其他研究使用其他的元資料，像是發言者的發文總數、該發言的回覆數量、每個追蹤該發言者的人的回應數量平均、發言者地理位置等，不過上述這些元資料對於仇恨言論偵測目前都沒有顯著幫助

(Waseem & Hovy, 2016; Zhong et al., 2016)。(Zhong et al., 2016)也指出，在他們的研究中，元資料對於他們的偵測沒有太大幫助的原因，可能是因為他們研究收集的是知名人物的帳號。這表示，雖然有些元資料對於仇恨言論偵測有幫助，但還是取決於原資料的類型與來源，用於知名人士的社群媒體言論上與普通用戶的會有差別。

2.2.8 非文字類仇恨言論

現今網路中，不只有文字訊息，也存在圖片、影片甚至聲音的訊息，這些非文字類的訊息也常常存在著仇恨言論，不過，(Schmidt & Wiegand, 2017)提到盡管視覺訊息對於仇恨言論有很大的影響，但在過去的仇恨言論偵測研究中，利用非文字類訊息做為特徵的數量卻很少。(Hosseinmardi et al., 2015)的研究便是以圖像標籤、分享內容、圖像標籤類別為特徵，(Zhong et al., 2016)的研究則利用圖像像素為特徵，並指出，將這些視覺特徵與從圖片文字取得的特徵結合，會得到最好的成效，也利用這些特徵來預測哪些圖片較容易遭受霸凌（bully-prone）、更容易吸引到霸凌言論，將之稱作霸凌觸發器（bullying triggers）。

2.3 角色

除了仇恨言論偵測以外，有些研究是針對霸凌事件中的角色，(Xu et al., 2012)的研究自動從整個霸凌事件中，指派角色給 Twitter 文章中的發文者以及被提及的人，其中角色包含霸凌者（bully）、受害者（victim）、協助者（assistant）、捍衛者（defender）、旁觀者（bystander）、加強者（reinforce）、回報者（reporter）和指責人（accuser）。(Sood et al., 2012)的研究則是自動預測該言論

告知的對象為前一個留言、或是針對其他人。(Silva, Mondal, Correa, Benevenuto, & Weber, 2016)的研究則是分析兩大社群媒體 Facebook 和 Whisper 中哪些目標最容易成為霸凌針對的對象，該研究結果發現，兩大平台前六種目標都是相同的，分別是種族、行為、身體特徵、性取向、階級、性別，而該研究為仇恨言論目標總共人工分了九個類別，和一個不屬於任何分類的其他，並為這些類別列出了其中幾個目標作為範例（如表 5）。

表 5 仇恨言論類別和範例目標(Silva et al., 2016)

類別	仇恨言論目標
種族	nigga, black people, white people
行為	insecure people, sensitive people
身體特徵	obese people, beautiful people
性取向	gay people, straight people
階級	ghetto people, rich people
性別	pregnant people, cunt, sexist people
文化	Chinese people, Indian people, Paki
殘疾	retard, bipolar people
宗教	religious people, jewish people
其他	drunk people, shallow people

2.4 預測社會事件

除了單一仇恨言論偵測研究外，有些研究也針對一段時間範圍內，極端負面的言論佔總體言論的比例，可以依此觀察社會大眾或個人情緒的變化。如果短時間內，仇恨言論數量、比例增加，可能顯示出發生了某些事情，也可以依此，在種族暴力、恐怖攻擊或其他犯罪發生前得知，進而做到預期治理（anticipatory governance）(Schmidt & Wiegand, 2017)。(Wang, Gerber, & Brown, 2012)的研究從

Twitter 資料中預測肇事逃逸的犯罪，試圖改善過去使用犯罪歷史、地理位置來預測肇事逃逸犯罪的情況。(Pete Burnap et al., 2015)的研究建立社群媒體張力自動檢測，該研究使用情感分析和特定主題的污辱、辱罵性詞彙資源，並將其可視化，也得到不錯的成效。

2.5 分類方法（Classification Methods）

仇恨言論偵測的方法大多是使用監督式學習，過去的研究中，大多數是使用支援向量機（Support Vector Machine, SVM）之類相較簡單的分類器，(Zhang & Luo, 2018)的研究即使用了 SVM 針對公開之仇恨言論資料集進行仇恨言論偵測，不過，當仇恨言論具有不同種的含義或含義為隱藏性時，僅依靠詞彙為特徵是不足的(Davidson et al., 2017; Dinakar et al., 2012; Kwok & Wang, 2013)。而近期也開始有研究使用深度學習的方式，(Mehdad & Tetreault, 2016)使用 Recurrent Neural Network Language Models（RNNLM）與深度學習進行仇恨言論偵測。(Badjatiya, Gupta, Gupta, & Varma, 2017)還有和(Gambäck & Sikdar, 2017)分別使用時間循環神經網路 Recurrent （Neural Networks, RNN）和卷積神經網路（Convolutional Neural Networks, CNN）進行社群媒體 Twitter 上的推文仇恨言論偵測，(Zhang & Luo, 2018)的研究除了使用 SVM，也同樣使用了卷積神經網路與多種卷積神經網路之變形進行仇恨言論偵測。而也有使用半監督式學習方式的研究，尤其是使用了自助法（bootstrapping）的研究，自助法用於仇恨言論偵測的目的不太一樣，一方面他可以用來獲得額外的訓練資料，如(Xiang et al., 2012)的研究，先將一批 Twitter 用戶依據他們文章內攻擊性用語的比例，分成優良使用者、不良使用者，再將不良使用者的所有言論都列為仇恨言論，加入至訓練集中。另一方面，自助

法也可以用於建立仇恨言論偵測使用的詞彙資源，(Gitari et al., 2015)的研究便使用此方法，增加研究的仇恨動詞辭典，以一個小的仇恨動詞清單為種子，並且依WordNet 關係，慢慢增加他們的同義詞與上位詞（hypernyms）。

而(Zampieri et al., 2019)針對社群媒體中攻擊性言論（Offensive Posts）的偵測研究中，雙向長短期記憶模型（Bidirectional LSTM, Bi-LSTM）與卷積神經網路的成效是較優於支援向量機。而(Pavlopoulos, Malakasiotis, & Androutsopoulos, 2017)提出了 RNN 模型結合自注意力機制（self-attention mechanism），用於辱罵留言（abusive comment）的偵測，而(MacAvaney et al., 2019)的研究則使用了BERT 語言模型與公開之仇恨言論資料集進行仇恨言論偵測。

雙向長短期記憶模型（Bidirectional LSTM, Bi-LSTM）

在提到 Bi-LSTM 前，必須先提到長短期記憶模型（Long Short-Term Memory, LSTM），由(Hochreiter & Schmidhuber, 1997)提出的，為 RNN 的一種，適合用於預測時間序列資料（sequential data）的模型，文字資料便為其一，主要是用於解決 RNN 其因可以學習的時間長度序列沒有限制，隨著輸入越來越多，產生長期依賴關係，而對前面的輸入感知力下降，發生梯度消失（gradient vanishing）或梯度爆炸（gradient exploding）的問題，也就是說，LSTM 對於較長的序列能有較 RNN 更好的表現，而 LSTM 與 RNN 的差別在於多了一個隱藏狀態c_t，稱為細胞狀態（cell state），用來記錄長期的訊息，通常c_t的改變很慢，都是由上一個狀態c_{t-1}傳過來的狀態再加上一些數值，而隱藏狀態h_t在每個不同的節點都會有很大的不同，LSTM 透過閥門的機制去限制訊息的傳輸量，分別是遺忘閥（forget gate）、輸入閥（input gate）、輸出閥（output gate），其數學公式所如下。

輸入閥：$i_t = \text{sigm}(W_{xi}\, x_t + W_{hi}\, h_{t-1})$

遺忘閥：$f_t = \text{sigm}\big(W_{xf}\, x_t + W_{hf}\, h_{t-1}\big)$

輸出閥：$o_t = \text{sigm}(W_{xo}\, x_t + W_{ho}\, h_{t-1})$

細胞狀態更新：$\tilde{c}_t = \tanh(W_{xi}\, x_t + W_{hc}\, h_{t-1})$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

LSTM 輸出：$h_t = o_t \odot \tanh(c_t)$

其中⊙代表逐元素相乘，sigm 代表 sigmoid 函數，輸出的範圍為 0~1。

$$\text{sigm}(z) = \frac{1}{1 + exp(-z)}$$

而h_{t-1}, h_t表示前一個隱藏狀態跟當下的隱藏狀態，W_i, W_o, W_f, W_c則表示輸入閥、輸出閥、遺忘閥和當下輸入狀態的權重矩陣。而這三個閥門的計算方式皆相同。

輸出閥o_t：目的為從細胞狀態c_{t-1}中產生隱藏狀態h_{t-1}，判斷c_{t-1}中哪些部份對於h_{t-1}是有用的，哪些是沒有用的。

輸入閥i_t：目的為將當下輸入的詞x_t的訊息加進c_t，判斷x_t對於整個句子的重要性，若i_t打開時，神經網路將不考慮x_t。

遺忘閥f_t：則判斷前一個細胞狀態c_{t-1}對於當下細胞狀態c_t的影響，因為當下輸入的詞x_t可能為接續前文的詞，也可能為新句子的開頭詞，所以f_t打開時。神經網路將不考慮前一個細胞狀態c_{t-1}。

細胞狀態c_t：綜合了當下的詞x_t和前一個細胞狀態c_{t-1}的訊息，當f_t關閉時，c_t可以不受參數的影響將梯度傳遞給c_{t-1}，這也是 LSTM 改善梯度消問題的關鍵。

不過，標準的單向 LSTM，只能由左至右接收詞向量與傳遞、產生隱藏狀態，導致模型較重視後來的輸入，只能利用到當前的詞之前的訊息，之後的訊息無法得知，為改善此問題，出現了雙向 LSTM（Bi-LSTM）如圖 4，由一個正向 LSTM（由左至右）與反向 LSTM（由右至左）結合，將正向與反向之結果結合，便可充分利用詞的上下文訊息，獲得文本雙向的語義關係。

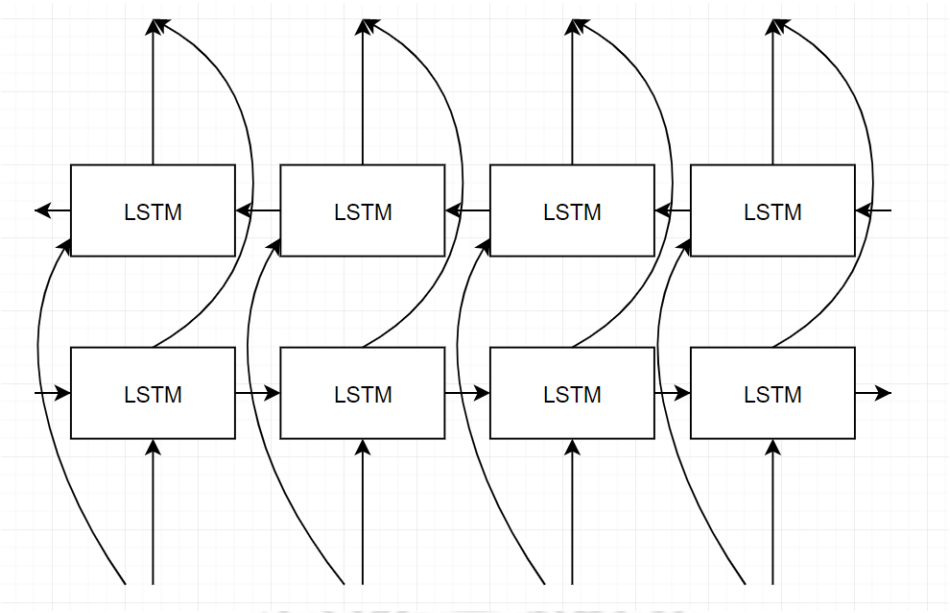


圖 4 Bi-LSTM model

BERT 語言模型

BERT（Bidirectional Encoder Representations from Transformers）模型為為 Google 團隊於 2018 年時所提出的語言模型，BERT 以(Vaswani et al., 2017)提出的 Transformer Encoder 為架構，以大量未標記的文本、透過非監督式學習進行模型的預訓練，再以此為基礎進行微調（fine tune）多個下游任務。

BERT 的預訓練是使用了兩個不同的任務，分別為遮蓋語言模型任務（Masked LM, Language Model）、下句話預測（Next Sentence Prediction）。BERT 使用了 Masked LM 以完成雙向語言模型作為第一個預訓練任務，做法為隨機將

語料中 15%的詞遮蔽，透過前後文的詞，預測這 15%被遮蔽的詞，以此獲取語料雙向的關係。而在這 15%詞彙中，遮蔽的方式如下。

其中，有 80%的詞彙是以“[MASK]”來遮蔽。

My car is white. -> My car is [MASK].

10%的詞彙是以任意的詞彙來代替。

My car is white. -> My car is sugar.

10%的詞彙是不做任何改變的。

My car is white. -> My car is white.

自然語言處理的任務中，常常會依賴句子與句子間的關係，像是問與答的任務，而語言模型往往沒辦法獲得這類關係，因此 BERT 使用了 Next Sentence Prediction 做為第二個預訓練任務，使用了語料中的兩句話做為訓練語料，預測第二句話是否為第一句話的下一句話，而在此預訓練任務中。也並不是所有的句子都是這樣處理的，有 50%的第二句話為正確的第一句話的下一句，而另外 50%則為語料中隨機的一句話。

而 BERT 的輸入，是由三個 embedding 的和（如下圖），分別為：

1. Token embedding 為當前詞的 embedding
2. Segment embedding 為當前詞所在句子的 index embedding
3. Position embedding 為當前詞所在位置的 index embedding

為了表示句子與句子間的區隔，以 segment embedding 和[SEP]做為區分，而句子的第一個 token，則有特殊含義，像是分類問題中的類別[CLS]。上述三個 embedding 的總和即為輸入的向量，如圖 5。

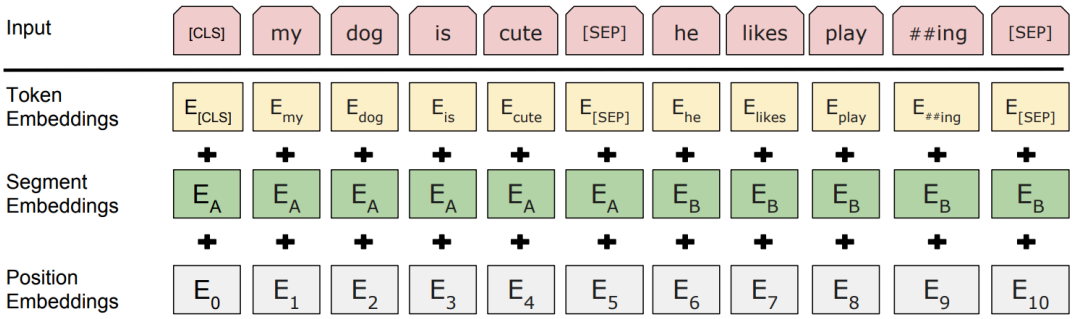


圖 5 BERT 語言模型之輸入(Devlin, Chang, Lee, & Toutanova, 2018)

以上述兩個任務所生成的預訓練模型結果，再針對不同的下游任務需求，設置單一輸出層，對模型進行 fine tune 微調。

2.6 評估方式

過去的仇恨言論研究中，通常都使用混淆矩陣（Confusion Matrix）中標準的精準率（precision）、召回率（recall）和 F1-score。精準度用於衡量系統預測為仇恨言論的數量中真實為仇恨言論（true positive）的百分比，召回率用於衡量我們期待系統預測出的仇恨言論（真實的有效值（ground truth））數量中，被預測為仇恨言論（true positive）的百分比。而 F1-score 則是兩者的調和平均數。（如圖 6）

	實際 YES	實際 NO
預測 YES	TP（true positive）	FP（false positive）
預測 NO	FN（false negative）	TN（true negative）

精準率 ＝TP /（TP+FP）

召回率 ＝TP /（TP+FN）

$$\text{F1-score} = \frac{2}{1/\text{精準率} + 1/\text{召回率}}$$

圖 6 混淆矩陣

也有些仇恨言論的研究使用微平均（micro-average）的精準率、召回率和 F1-score (Badjatiya et al., 2017; Gambäck & Sikdar, 2017; Park & Fung, 2017; Waseem, 2016; Waseem & Hovy, 2016; Yuan et al., 2016)，因為當使用不平衡資料集，數量較多的類別（dominant class）的數量遠多於數量較少的類別（minority classes）時，微平均可以掩蓋數量較少的類別其真實的成效，因為即使數量較少的類別擁有的 F1-score 遠高於或遠低於數量較多的類別，也不會對整個資料集的 micro-F1 值造成什麼影響。而仇恨言論就是這種類別數量差異極大的不平衡資料集，仇恨言論數量佔整個資料集的比例極低。

微平均的精準率、召回率和 F1-score 三者皆為相同數值，是因為在此時 TP、FP 和 FN 的算法皆與上述圖 6 的算法不同，三者的計算方式如下，為整個樣本一起計算，類別不會造成影響，因此也能對於類別數量大於二的分類工作進行評估，像是一些仇恨言論偵測的研究(Van Hee et al., 2015)將資料集分為正面、負面、中性三種類別。

TP：為整個樣本中被預測且正確的數量（預測 = 實際）。

FP：為整個樣本中預測的值與實際的值不同，預測錯誤的數量。

FN：為整個樣本中預測的值與實際的值不同，實際的值的數量。

由上述計算可得知，FP 若是增加，FN 也會增加相同的值，反之亦然，因此 FP 與 FN 必定會是相同的值，這也是為什麼微平均的精準率、召回率和 F1 Score 三者皆為相同數值。

2.7 資料集

對於仇恨言論偵測而言，有標記好的語料庫作為資料集是很重要的，而有許多研究都自行收集語料進行標記，像是(Pete Burnap et al., 2015; Peter Burnap & Williams, 2014; Pete Burnap & Williams, 2015; Silva et al., 2016; Xiang et al., 2012; Xu et al., 2012) 使用 twitter 資料，(Hosseinmardi et al., 2015; Zhong et al., 2016)使用 Instagram 資料，(Djuric et al., 2015; Nobata et al., 2016; Warner & Hirschberg, 2012)使用 Yahoo! 資料，(Dinakar et al., 2012)使用 YouTube 資料，(Van Hee et al., 2015)使用 ask.fm 資料，(Dinakar et al., 2012) 使用 Formspring 資料，(Razavi et al., 2010)使用 Usenet 資料，(Silva et al., 2016)使用 Whisper 資料，(Chau & Xu, 2007)使用 Xanga 資料。

(Fortuna & Nunes, 2018) 的研究整理了 2016 年 9 月 1 日至 2017 年 5 月 18 日間共 128 篇與仇恨言論相關的研究，為這幾篇研究分為「法律與社會科學」、「電腦科學與工程」兩類，並整理出「電腦科學與工程」類別使用了哪些社群媒體的資料（如表 6）。

表 6 「電腦科學與工程」類別使用的社群媒體(Fortuna & Nunes, 2018)

社群媒體	使用次數
Twitter	16
Sites	5
YouTube	3
Yahoo! finance	2
American Jewish Congress（AJC）sites	1
Ask.fm	1
Blogs	1
Documents	1
Facebook	1
formspring.me	1
myspace.com	1
Tumblr	1
Whisper	1
White supremacist forums	1
Yahoo news	1
Yahoo!	1

從上述可知，最多研究使用的資料來源是 Twitter。而這些研究使用不同網站的資料，會因網站的類型與目的不同，可能會各自產生不同的仇恨言論特徵(Schmidt & Wiegand, 2017)，而語料庫使用的大小也不一，(Dinakar et al., 2012)的研究使用大約 100 條標籤過的留言與知識庫特徵結合。(Van Hee et al., 2015)和(Djuric et al., 2015)則是使用數千條留言。而(Schmidt & Wiegand, 2017)提到，在做仇恨言論偵測的有效性評估時，必須要考慮到資料集的大小，該評估結果可能只是對於特定大小的數據的結果，若用與研究使用的語料庫不同大小的資料，可能會有不同的結果。

(Fortuna & Nunes, 2018)的研究整理出一些資料集跟文本（如表 7）並提到，雖然現在網路上有許多關於仇恨言論的資料集跟文本，但使用哪一個比較好都還沒有一個定論。

表 7 用於仇恨言論偵測的資料集與文本(Fortuna & Nunes, 2018)

名稱	分佈	年份	類型	數量	使用類別	使用語言
Hate Speech Twitter annotations[9]	GitHub repository	2016	資料集	16914	性別、種族	英文
Hate Speech identification[10]	公開的	2015	資料集	14510	冒犯性仇恨言論與非仇恨言論、非冒犯性言論	英文
Abusive language dataset[11]	不公開	2016	資料集	2000	仇恨言論、非冒犯性言論	英文
German Hatespeech Refugees[12]	創用 CC3.0 授權	2016	資料集	470	仇恨言論、非冒犯性言論	德文
Hatebase[13]	公開的	2017	文本	-	-	多語言
Hades[14]	公開的	2016	文本	-	-	荷蘭語
Hate Speech and offensive language[15]	公開的	2017	文本	-	-	英文

[9] Waseem, Z. （2016）. Hate speech Twitter annotations. Retrieved from https://github.com/ZeerakW/hatespeech.

[10] CrowdFlower. 2017. Data for everyone. Retrieved from https://www.crowdflower.com/data-for-everyone/.

[11] Yahoo! 2017. Webscope datasets. Retrieved from https://webscope.sandbox.yahoo. com/.

[12] UCSM. 2016. IWG hatespeech public. Retrieved from https://github.com/UCSM-DUE/.

[13] Kaggle. 2013. Detecting insults in social commentary. Retrieved from https://www.kaggle.com/c/detecting-insults-in-social-commentary/data.

[14] CLiPS. 2016. HADES. Retrieved from https://github.com/clips/hades.

[15] Davidson, T. （2017）. Automated hate speech detection and the problem of offensive language. Retrieved from https://github.com/t-davidson/hate-speech-and-offensive-language.

第三章 研究方法與系統架構

3.1 前言

第二章，我們探討了過去仇恨言論偵測相關研究，有許多不同的方法與對仇恨言論的定義，而本研究，將會使用神經網路模型做仇恨言論偵測的分類工作，本章節，將會介紹整個實驗的流程與細節。

3.2 系統架構與流程

本研究所提出深度學習與自然語言處理於仇恨言論之自動偵測研究的系統架構如圖 7 所示。

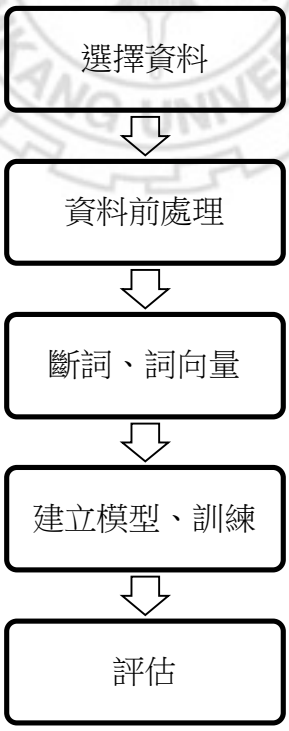


圖 7 深度學習與自然語言處理於仇恨言論之自動偵測研究

1. 選擇資料：選擇仇恨言論資料集。

2. 資料前處理： 將資料集內語料進行整理，剔除不必要之欄位、符號。

3. 斷詞、詞向量：將資料集內語料進行斷詞、將詞彙轉換為詞向量。

4. 建立模型、訓練：建立深度學習模型，以詞向量為輸入進行訓練。

5. 評估：以測試資料集評估訓練後的模型之正確率、精準率、召回率、F1-score。

3.3 資料集

本研究使用兩個不同的資料集語料做為深度學習模型的輸入，分別為 HatebaseTwitter 資料集與 3000_tweets_hate_goldlabel 資料集。

3.3.1 HatebaseTwitter 資料集:

HatebaseTwitter 資料集，由(Davidson et al., 2017)所搜集和標記，此資料集包含了 24803 條 Twitter 推文，為 csv 格式檔案，該研究建立資料及的過程如下，首先，他們發表了一個仇恨言論詞典 Hatebase，並依此詞典內的詞彙，從約 33000 名 Twitter 用戶搜集了約 8500 萬條推文，再從中隨機挑選了 25000 條推文，再透過群眾外包（crowdsourcing）交給美國 CrowdsFlower 公司，將這些推文進行標記，分為仇恨言論（hate speech）、令人反感的（offensive）言論（但不是仇恨言論）、其他（neither），令人反感的言論為像是某些發言者習慣性使用髒話、或是發言含有髒話只是為了情緒的表達，而並非仇恨言論。而當推文標記無法取得共識時，則刪除該推文。最後，該資料集一共有六個欄位（如圖 8），分別為：

「count」：CrowdsFlower 註釋者投票數量，最小值設定為 3。

「hate_speech」: CrowdsFlower 註釋者認為為仇恨言論之數量。

「offensive_language」：CrowdsFlower 註釋者認為為令人反感言論之數量。

「neither」：CrowdsFlower 註釋者認為為其他之數量。

「class」：CrowdsFlower 註釋者最終投票結果，仇恨言論為 0、令人反感的言論為 1、其他為 2。

「tweet」：為 Twitter 推文內容。

count	hate_speec	offensive_l	neither	class	tweet
3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you sh
3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confu
3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who t
3	1	2	0	1	!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fuck
3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
3	0	3	0	1	!!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!”
3	0	3	0	1	" & you might not get ya bitch back & thats that "
3	1	2	0	1	" @rhythmixx_ :hobbies include: fighting Mariam" bitch
3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh
3	0	3	0	1	" Murda Gang bitch its Gang Land "
3	0	2	1	1	" So hoes that smoke are losers ? " yea ... go on IG
3	0	3	0	1	" bad bitches is the only thing that i like "
3	1	2	0	1	" bitch get up off me "
3	0	3	0	1	" bitch nigga miss me with it "
3	0	3	0	1	" bitch plz whatever "
3	1	2	0	1	" bitch who do you love "
3	0	3	0	1	" bitches get cut off everyday B "
3	0	3	0	1	" black bottle & a bad bitch "

圖 8 HatebaseTwitter 部分資料集

3.3.1 3000_tweets_hate_goldlabel 資料集:

3000_tweets_hate_goldlabel 資料集是由(Teh & Cheng, 2019)所搜集和標記，此資料集包含了 3000 條 Twitter 推文，為 csv 格式檔案，為發文地點來自馬來西亞、美國、澳洲，這三個地點各 1000 條推文，該研究建立資料集的過程如

下，首先，使用了一個名為 Twitter Archiver 的 google 插件，依照 Twitter 上使用頻率最高的前 20 個辱罵詞彙檢索，在 2017 年 9 月至 2018 年 5 月之間，從 Twitter 收集了 26250 則推文，其中 17661 則來自澳洲的用戶，4435 則來自美國的用戶和 4154 則來自馬來西亞的推文。為了消除各地方言的出現，設定了定理位置的限制，為首都 10,000 英里以內的大都市地區，並且在搜集到之後，將除英文外還混雜其他語言之推文剔除，最後再從中各取 1000 條推文。該資料集共有十個欄位，分別為:

「Date」:該 Twitter 推文發文時間。

「Screen Name」: 該發文者顯示之名稱。

「Full Name」:該發文者之全名。

「Tweet Text」: 該 Twitter 推文。

「Location」: 發文者之地理位置。

「R#1」、「R#2」、「R#3」:判斷是否為仇恨言論之投票。

「Label」:若「R#1」、「R#2」、「R#3」有達到兩票，即標記為仇恨言論。

還有一個欄位是該 Twitter 推文之分類。該資料集如圖 9。

Date	Screen Nam	Full Name	Tweet Text	Location		R#1	R#2	R#3	Label
11/1/2017 8:12:23	@aetheriun	move im gay	I forgot the outer suburbs were shit and I'm wearing s	Melbourne, Victoria	Racial	y	y		y
11/1/2017 8:17:26	@dale_fitzhe	Dale Fitzhenry	@BrendanSchaub Damn you americans like some shi	Melbourne, Victoria	Ethnicity		y	y	y
11/1/2017 8:19:02	@nushyb	AnushkaB	*whispers to self* holy hell how is it November?	Melbourne, Victoria	Religion	y	y		y
11/1/2017 8:20:32	@CatherineD	Catherine Deveny	Fuck reading. Make this the summer of writing. Book	Melbourne. My hometow	Other				
11/1/2017 8:20:52	@lildmg	Dani	I'm a caffeine slut. I'll take it anywhere, anytime, and	Melbourne, Australia	Behaviour				
11/1/2017 8:31:59	@liv_croagh	Liv Croagh	@nickschadegg What the hell si that meant to mean!?	Melbourne	Other				
11/1/2017 8:32:19	@BASEDJU	「basedjuanempire」	im gonna be on instagram live in a minute talking shit	melbourne, australia	Other				
11/1/2017 8:36:32	@timothy_ne	Sure, Tim Newport	RT @text_publishing: Claire Vaye Watkins' list of 'W	Melbourne, Australia	Religion				
11/1/2017 8:37:33	@afcoory	Anne Frandi-Coory	RT @JohnWren1950: There is a demarcation dispute	Melbourne	Physical		y		
11/1/2017 8:57:02	@IzzyMelhe	Isabel	@DarkWhite292 fuck yeah on my way	Melbourne, Victoria	Sexual Ori		y		
11/1/2017 8:58:18	@LarryShort	Larry Short	@SkyNewsAust @JulieBishopMP Leaving it to the P	Melbourne, Victoria	Racial				
11/1/2017 8:58:25	@NotUnderE	Barbara Roberts	@AlsoACarpenter @DavidBancz @RScottClark @L	Melbourne, Australia	Other			y	
11/1/2017 9:16:15	@justjoshuaa	🍆♂	@cassietee_ I'd suck the dick of the dude in that phot	Melbourne	Gender	y	y	y	y
11/1/2017 9:20:16	@changeofhr	aimee🍊	I GOT TICKETS TO DUA LIPA FUCK YES	Melbourne	Sexual Ori				
11/1/2017 9:28:13	@theboyofch	Benjamin scrEEm 💀🦃	@JennRavenna @CarmenSinek @andrewkmar Oh yi	Melbourne, Victoria	Sexual Ori			y	
11/1/2017 9:34:00	@LiadeleH	Liadele	Haters will say my kids DIY makeup for Halloween is	Melbourne, Australia	Other				
11/1/2017 9:40:34	@JohnicAdv	JohnicAdventure	@SuperSonic512Tx @Rhymestyle God damn it.	Melbourne	Religion				
11/1/2017 9:41:40	@kasemc	Kase M.	2017 has been shit and there's still two months to get	Melbourne, Australia	Religion				
11/1/2017 9:49:10	@amyiox_	minhyukkie🐶	At this point in time, I am a lost cause. Because I don'	Melbourne, Victoria	Religion			y	
11/1/2017 9:50:06	@swxxii	Simon	"You know what would make shit smell better? Vanil	Melbourne, AU	Other				
11/1/2017 9:50:47	@AForeigne	AGAIN, REGULAR GUY	RT @wheelswordsmith: [me, a pilot] this is your capt	ayy vee el	Other				

圖 9 3000_tweets_hate_goldlabel 部分資料集

3.4 資料前處理

因為資料集中仍然有許多我們不會使用到的資料，像是 HatebaseTwitter 資料集進行群眾外包標記時對每則推文的投票結果，本研究僅使用到 ”class”、“Tweet” 兩個欄位，須先將其他不必要欄位剔除，而 3000_tweets_hate_goldlabel 資料集本研究僅使用” Tweet Text”、“Label” 兩個欄位。由於之後需進行詞向量工作，先將“Tweet”欄位的推文內的標點符號、其他特殊符號以及換行符號剔除，下列為剔除之項目:

*“ ! " # $ % & （ ） * +, - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~ \t \n ”*

而 HatebaseTwitter 資料集分為三類，本研究除了使用原始資料集進行三分類任務預測外，也將該資料集中令人反感的言論與其他兩類合併為一類，標記為 1，總共分為仇恨言論、其他，進行兩分類預測。而 3000_tweets_hate_gold-label 資料集本研究除了使用原始資料集三個國家的子資料集，也將三個子資料集合併為一個 3000 則推文的資料集，進行兩分類預測。

3.5 詞向量

如第二章 2.2.2 所述，本研究也將資料集中的語料單詞，轉換為詞向量，作為神經網路模型之輸入，我們可以以詞彙之間的詞向量距離，判斷詞彙意思之間的關係，詞義相近的字會有較小的向量距離，反之亦然如圖 10。如此，神經網路模型在訓練時，便可得知詞彙在語句中與前後文之關係，提高訓練成效。

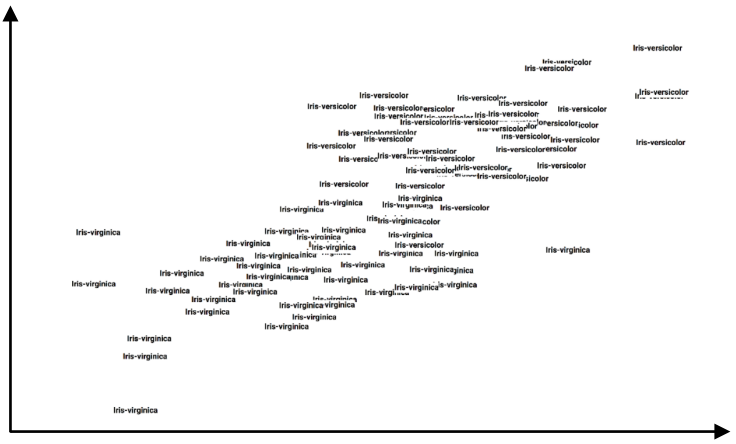


圖 10 詞向量空間可視化

3.6 建立深度學習模型

本研究使用深度學習中的 BERT 語言模型與雙向長短期記憶模型（Bi-LSTM）對 3.3 章所述之兩個資料集進行仇恨言論預測。

1. BERT 語言模型

首先，我們使用了 google 團隊提出的語言模型 BERT，使用的預訓練模型版本為 BERT-Base, Uncased，與 Large 版本相比，為層數較少的版本，使用了 12 層的 Transformer，隱藏層維度為 768，multi-head Attention 的參數為 12，google 團隊使用了英文維基百科文本與 BookCorpus 作為此預訓練模型的訓練集，前者包含了兩億五千萬個詞，後者則包含了八千萬個詞。

確定了 BERT 預訓練模型版本之後，我們將上述兩個資料集分別分成訓練集與測試集，訓練集作為微調預訓練模型訓練之輸入，微調訓練完成之模型，再以測試集之資料輸入評估模型，流程如圖 11。

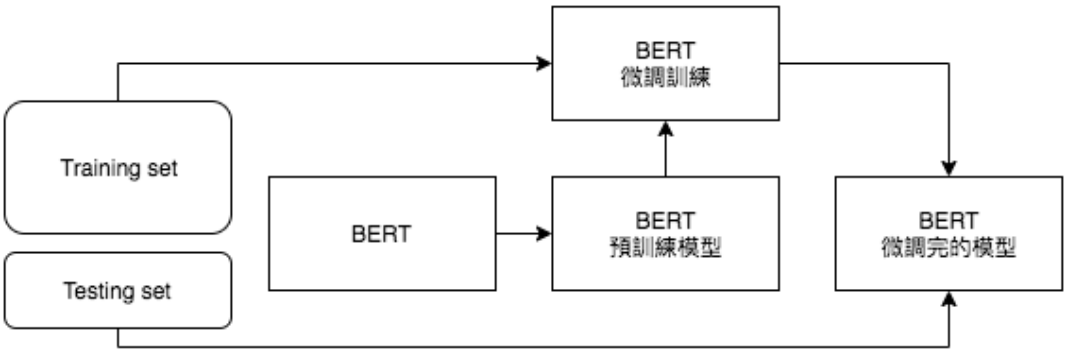


圖 11 BERT 語言模型流程圖

2. 雙向長短期記憶模型（Bi-LSTM）

第二個使用的模型則為雙向長短期記憶模型，我們使用 Keras（http://keras.io/）作為建立該模型之工具，Keras 為一個開放原始碼的高階深度學習程式庫，並能夠運行於 Tensorflow 與 Theno 之上，可以使用 Keras 建立深度學習模型，進行訓練，並且能夠透過 GPU 資源加速運算的速度、評估模型準確率、進行預測。同樣的，我們將上述兩個資料集分別分成訓練集與測試集，分別作為模型訓練與評估之輸入。

而本研究實驗流程分為三個步驟：處理資料集文字、深度學習模型訓練、深度學習模型評估。

3.6.1 處理資料集文字

如前述 3.4 章節，我們獲得了沒有標點符號與特殊符號的資料集後，我們需將句子進行斷詞處理，而在自然語言處理中，英文的斷詞方式相較簡單，為將詞彙依句子中的空格分開。由於深度學習模型只能接受數字，我們必須將資料集中

的文字轉換為數字，為此，在斷詞後，我們須為這些文字建立一組字典，在此，我們使用了 Keras 提供的 Tokenizer 模組，建立 token 字典的方式如下：

（1） 設定字典的字數為 3000。

（2） 讀取訓練集中的詞彙，並依照辭彙出現的頻率高低進行排序，將排序前 3000 的詞彙加入字典中。

（3） 字典中顯示為詞彙與排序數字。

如此，我們便得到了該資料集中常用的 3000 個字的字典與其不重複的編號。在這之後，我們須利用此 token 字典，將自然語言句子，轉換為數字序列，即是使用此 token 字典中不重複的編號，取代自然語言詞彙。然而，每一個句子長短不一，由於之後要進行深度學習模型訓練，我們必須將數字序列的長度固定，在此，我們設定數字序列長度統一為 150，若是數字序列長度為 70，我們便在序列前補上 80 個 0，若是長度為 175，我們便將序列從後面刪除 25 個數子，如此一來，我們便能將長度均為 150 的數字序列輸入進嵌入層（embedding layer）進行詞向量的工作，在句子數字序列時，數字與詞彙意思沒有任何相關，再轉換為詞向量後意思接近的詞彙，會有距離較為接近的向量。上述流程如下圖 12。

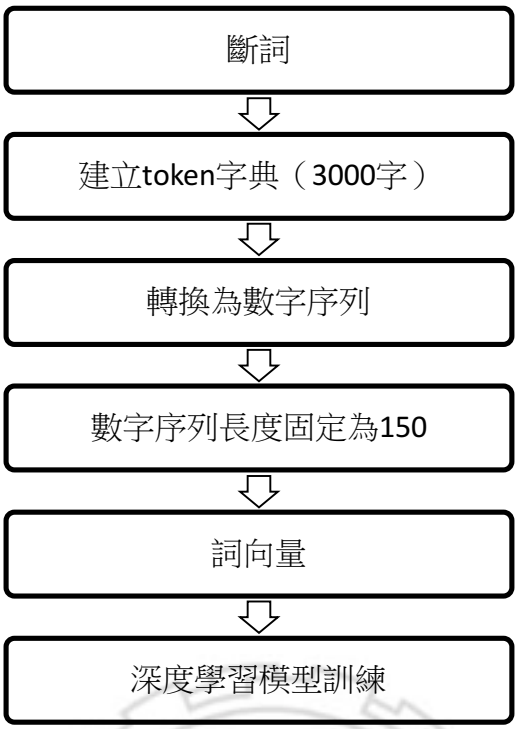


圖 12 處理資料集文字流程圖

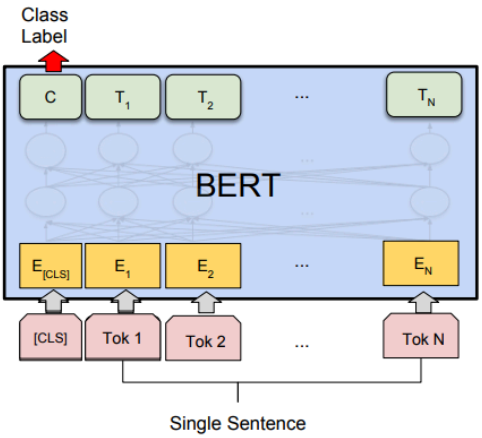
上述流程是針對雙向長短期記憶模型所做的處理，而 BERT 模型中，含有另一種對自然語言處理的方式 WordPiece，則是將單一詞彙做分割為片段，WordPiece 主要的作法為 BPE（Byte-Pair Encoding）雙字節編碼，即是將單一詞彙再拆開，例如“walked”、“walking”、“walks”，這三個英語詞彙，意思皆為走路，但若是我們使用傳統的詞向量方式，會將三者認為不同的詞彙，在英語中，由於不同的狀態，會導致相似的辭意有不同的詞尾，這將會導致詞彙表的數量變大，BPE 算法則會將上述三個詞彙拆分為“walk”、“ed”、“ing”、“s”，以此有效減少詞彙表的數量。

3.6.2 深度學習模型訓練

本章節分別講解本研究使用的兩個模型，BERT 模型與雙向長短期記憶模型之模型訓練細節。

1. BERT 預訓練模型微調

確定了訓練資料與 BERT 預訓練模型版本之後，我們便要將 BERT 預訓練模型進行微調，(Devlin et al., 2018)在 BERT 論文中有提出了四種常見的微調任務情境，而本研究使用的便是其中之一，單句子輸入的分類任務。首先，我們將訓練資料，分為訓練及與測試集以及驗證集，並改寫成與 BERT 輸入相同的格式 tsv 檔案，接著便依照我們的實驗需求，加入一層輸出層成為下游任務模型，並且訓練，驗證集於訓練時使用，測試集則為訓練完成後做預測、評估模型。我們需要建立一個新的 DataProcesser，目的為讀取需要的資料集，並依照分類任務進行設定，本研究為二分類，仇恨言論與否（0 與 1），而後將輸入的語料轉換為三種不同的向量進行加總，作為模型輸入，分別為 Token embedding、Segment embedding、Position embedding。圖 13 為本研究使用之單句分類任務微調圖，將句子插入第一個 token“[CLS]”，作為是否為仇恨言論的分類符號，E 表示嵌入的輸入，Ti 為第 i 個標記的上下文表示。最終，模型將輸出兩個類別的機率，如 [0.9122405, 0.08775952]，前者為 0 之機率，後者為 1 之機率。而本研究 BERT 微調時參數設定如下，max length sequence（句子長度）設定為 100，train batch size 設定為 32，learning rate 設定為 2e-5。



(b) Single Sentence Classification Tasks:
SST-2, CoLA

圖 13 BERT 微調情境(Devlin et al., 2018)

2. 雙向長短期記憶模型（Bi-LSTM）

本研究使用 Keras 建立 Bi-LSTM，神經網路模型架構如圖 14，為防止過度擬和，在第二層嵌入層之後新增 Drop out 層，設定為 0.3，第四層為 Bi-LSTM 層，在此層內的 Drop out 部分設定為 0.3，recurrent dropout 部分為 0.2，第五層為平坦層（Flatten），功能為將原本多維度的輸入一維化。第六層為全連接層（Dense），設定輸出 unit 為 256，激活函數為 sigmoid，而後為 Drop out 層，設定為 0.3，最後則是輸出，全連接層，設定輸出 unit 為 1，激活函數為 sigmoid。損失函數（Loss Function）使用二值交叉熵（Binary Cross Entropy）。優化器（optimizer）使用 adam。

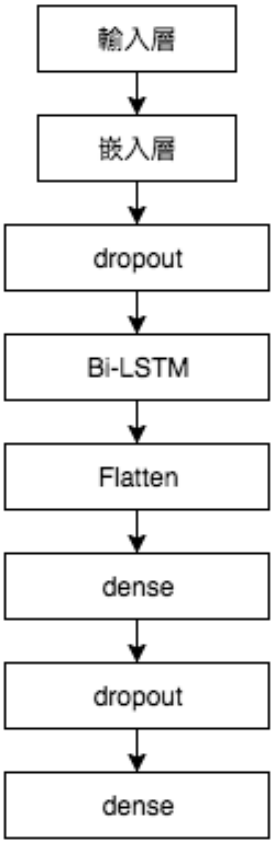


圖 14 Bi-LSTM 模型架構

3.6.3 深度學習模型評估

完成深度學習的訓練後，我們必須要有評估模型優劣的標準，而由於仇恨言論資料集皆為分布不均的極度不平衡資料集，我們除了正確率外，還必須考慮其他評估方式，在此我們使用精準率、召回率和 F1 score 進行模型評估，精準率是以預測的角度進行評估，以預測為仇恨言論之推文中，有多少推文真實為仇恨言論，準確率為何，而召回率則是以真實的角度進行評估，以實際為仇恨言論之推文中，有多少推文被預測為仇恨言論，準確率為何，而 F1 score 則是兩者的調和平均數之綜合評估標準。

為了得到上述三個評估數值，我們將兩個資料集的各個模型預測訓練集的結

果以混淆矩陣的方式呈現，混淆矩陣的 X 軸為真實結果（True label），Y 軸為預測結果（Predicted label）。由於我們的實驗中含有二分類任務與三分類任務，兩者計算的方式又不相同，對於二分類而言如表 8，會有四個數值，預測正確的仇恨言論（True Positive）、預測錯誤的仇恨言論（False Negative）、預測正確的其他言論（True Negative）、預測錯誤的其他言論（False Positive），以此四個數值計算精準率、召回率和 F1 score。而對於三分類任務而言如表 9，一樣有四個數值，預測正確的仇恨言論、預測錯誤的仇恨言論（實際為令人反感的言論或其他）、預測正確的令人反感的言論或其他、預測錯誤的令人反感的言論或其他（實際為仇恨言論）。

表 8　2 分類之混淆矩陣

預測			
	hate	預測正確的仇恨言論（True Positive）	預測錯誤的其他言論（False Positive）
	no	預測錯誤的仇恨言論（False Negative）	預測正確的其他言論（True Negative）
		hate	no
		真實	

資料來源：本研究整理

表 9　3 分類之混淆矩陣

預測			
	hate	預測正確的仇恨言論	預測錯誤的其他言論（實際為令人反感的言論或其他）
	offensive or neither	預測錯誤的仇恨言論（實際為仇恨言論）	預測正確的令人反感的言論或其他
		hate	offensive or neither
		真實	

資料來源：本研究整理

由於本研究關注在於仇恨言論是否能正確被偵測出來，因此，我們計算精準率、召回率和 F1 score 時，並非計算整體之數值，而是針對仇恨言論進行計算。然後由於我們關注在於仇恨言論，因此二分類任務中「預測正確的其他言論」與三分類任務中「預測正確的令人反感的言論或其他」，這兩個 True Negative 欄位皆不會被考慮進去，因為這個欄位為最普遍的情況，也就是不為仇恨言論且被預測正確的欄位，數量也佔絕大多數，將之計算進去將會使得整體數值提升，但將會影響我們判別模型之優劣。而計算精準率、召回率和 F1 score 方式如下。

精準率 :
預測正確的仇恨言論 / (預測正確的仇恨言論+預測錯誤的其他言論)
召回率 :
預測正確的仇恨言論 / (預測正確的仇恨言論+預測錯誤的仇恨言論)
F1 score:
2 / (1 /精準率 ＋ 1 /召回率)

3.7 實驗環境

本研究實驗過程皆使用 python3 語言，而由於 BERT 模型所需之運算效能較大，針對兩個模型使用了不同的實驗環境。

1. BERT 語言模型實驗環境

系統環境為，作業系統: Ubuntu 18.04.1 LTS， CPU 為 Intel Core i7-8700K (3.70GHz, 6 cores) x 2，RAM 為 64GB， GPU 型號為 NVIDIA GeForce RTX 2080 Ti (1.53GHz, 4352 cores, 11GB DDR6 RAM) x 2。

2. 雙向長短期記憶模型（Bi-LSTM）實驗環境

於 Google Colab （Colaboratory）上進行編譯、運算，為 Google 提供的雲端 Jupyter Notebook 開發環境，系統環境為，作業系統: Ubuntu 18.04.1，提供的 CPU 為 Intel（R） Xeon（R）TwinCore @ 2.20GHz x 2，RAM 為 13GB，空間為 347GB，GPU 型號為 Tesla K80 with 4992 cores at 556MHz + 11GB Memory。

第四章 資料分析與實驗結果

本章節將詳細的呈現與討論本研究之實驗結果，將包含實驗使用之資料集之資料分配、實驗設定與說明、實驗結果與分析、各類結果之比較。

4.1 資料分配

本研究使用兩個仇恨言論資料集進行實驗，分別為 HatebaseTwitter 資料集和 3000_tweets_hate_goldlabel 資料集，皆以 8:2 分為訓練集與測試集。

4.1.1 HatebaseTwitter 資料集

HatebaseTwitter 為本研究使用的兩個實驗資料集之一，該資料集內含有 24783 則 twitter 推文，而其中，「仇恨言論」有 1430 則，「令人反感的言論」有 19190 則，「其他」則有 4163 則，共有 23353 則。本研究以此資料集進行三分類訓練，除此之外，再將「令人反感的言論」及「其他」合併為一類「其他」，分為兩類進行訓練。再以 8:2 分為訓練集 19784 則推文、測試集 4999 則推文，表 10、11 為將三分類資料集與二分類資料集分為訓練集、測試集，各自的數量與其中標記數量。

表 10 HatebaseTwitter 三分類資料集訓練集與測試集數量

	總數	仇恨言論	令人反感的言論	其他
訓練集	19784	1121	15252	3311
測試集	4999	309	3838	852

資料來源：本研究整理

表 11 HatebaseTwitter 兩分類資料集訓練集與測試集數量

	總數	仇恨言論	其他
訓練集	19784	1121	18663
測試集	4999	309	4690

資料來源：本研究整理

4.1.2 3000_tweets_hate_goldlabel 資料集

3000_tweets_hate_goldlabel 為本研究使用的兩個實驗資料集之一，該資料集內含有共 3000 則 twitter 推文，本研究先各別將馬來西亞、美國、澳洲各 1000 則單獨進行訓練，再將這三個國家內各自 1000 則推文合併進行訓練。而其中「仇恨言論」有 668 則，「其他」則有 2332 則。再以 8:2 分為訓練集、測試集，表 12、13、14、15 為三個國家資料集與合併後分為訓練集、測試集，各自的數量與其中標記數量。

表 12 3000_tweets_hate_goldlabel 馬來西亞資料集訓練集與測試集數量

	總數	仇恨言論	其他
訓練集	789	185	694
測試集	211	54	157

資料來源：本研究整理

表 13 3000_tweets_hate_goldlabel 美國資料集訓練集與測試集數量

	總數	仇恨言論	其他
訓練集	789	147	642
測試集	211	47	164

資料來源：本研究整理

表 14 3000_tweets_hate_goldlabel 澳洲資料集訓練集與測試集數量

	總數	仇恨言論	其他
訓練集	789	357	432
測試集	211	93	118

資料來源：本研究整理

表 15 3000_tweets_hate_goldlabel 全部資料集訓練集與測試集數量

	總數	仇恨言論	其他
訓練集	2391	529	1862
測試集	609	139	470

資料來源：本研究整理

本研究發現我們所使用的這兩個資料集如(Zhang & Luo, 2018)所述，仇恨言論資料集多為極端不平衡之資料集，仇恨言論佔資料集的比例較低，不過，為了能夠真實呈現網路實際的狀況，本研究還是以非平均採樣的資料進行實驗。

4.2 實驗設定與說明

如第三章所述，本研究將各別使用兩個資料集作為語料，以兩種不同的模型進行深度學習網路實驗，判斷 Twitter 推文是否為仇恨言論。

實驗一：使用 BERT 模型。

實驗二：使用 Bi-LSTM 模型。

而實驗結果將以模型預測測試集之混淆矩陣呈現，最終再以正確率、精準率、召回率、F1 Score 作為評估標準。

4.3 實驗結果與分析

如前述，分別呈現兩個資料集之不同分配方式進行實驗一與實驗二之結果，將同個資料集的兩個實驗結果進行比較與分析。

4.3.1 HatebaseTwitter 資料集

本章節為使用 HatebaseTwitter 資料集，如 4.1.1 章節所述，將資料集原始的三分類與將資料集合併為二分類，使用 BERT 模型與 Bi-LSTM 模型進行訓練預測之結果。

3 分類：

表 16 為 HatebaseTwitter 資料集原始三分類以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終 loss 值。BERT 模型訓練了 10 個 epoch，Bi-LSTM 模型訓練了 40 個 epoch。由此可以看出，於正確率的表現，使用 BERT 模型達到 0.90，是高於使用 Bi-LSTM 模型之 0.76 的。

而表 17 與表 18 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，由此我們也可以看出，由於原始三分類之資料集之資料大部分都是集中在令人反感的言論，因為資料集這樣的分布不均，導致使用 Bi-LSTM 將所有的推文預測為令人反感的言論，模型即使是將大部分 twitter 推文預測為非仇恨言論，也會得到相當高的正確率。而使用 BERT 模型表現則較優，精準率為 0.53、召回率為 0.44，而 F1 score 則為 0.48。

表 16 HatebaseTwitter 資料集 3 分類正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.90	0.91	0.26	0.26	490
Bi-LSTM	0.76	0.79	0.22	0.20	3687

資料來源：本研究整理

1. BERT

表 17 BERT、HatebaseTwitter 資料集 3 分類之混淆矩陣

預測				
	hate	137	100	19
	offensive	158	3662	90
	neither	14	76	743
		hate	offensive	neither

精準率 :0.53
召回率 :0.44
F1 score:0.48

真實

資料來源：本研究整理

2. Bi-LSTM

表 18 Bi-LSTM、HatebaseTwitter 資料集 3 分類之混淆矩陣

預測				
	hate	0	0	0
	offensive	309	3838	852
	neither	0	0	0
		hate	offensive	neither

精準率 : 0
召回率 : 0
F1 score: 0

真實

資料來源：本研究整理

2 分類：

表 19 為 HatebaseTwitter 資料集，將令人反感的言論與其他合併為一類，進行二分類任務，以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終 loss 值。BERT 模型訓練了 10 個 epoch，Bi-

LSTM 模型訓練了 10 個 epoch。在此二分類任務時，可以看出使用使用 BERT 模型與使用 Bi-LSTM 模型之測試集正確率表現與該資料集原始之三分類結果不同，二分類之測試集正確率皆為 0.93。

而表 20 與表 21 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，在此可以看出，即使於測試集之正確率兩者皆相同，不過於其他的評估標準來看，對於偵測仇恨言論的表現，使用 BERT 模型是較優於使用 Bi-LSTM 模型的，使用 BERT 模型的 F1 score 達到 0.44，而使用 Bi-LSTM 模型則僅有 0.23。

表 19 HatebaseTwitter 資料集 2 分類正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.93	0.94	0.16	0.16	490
Bi-LSTM	0.93	0.95	0.11	0.17	994

資料來源：本研究整理

1. BERT 模型

表 20 BERT、HatebaseTwitter 資料集 2 分類之混淆矩陣

預測				
	hate	125	121	精準率 :0.5
	no	184	4569	召回率 :0.4
		hate	no	F1 score:0.44
		真實		

資料來源：本研究整理

2. Bi-LSTM 模型

表 21 Bi-LSTM、HatebaseTwitter 資料集 2 分類之混淆矩陣

預測				
	hate	48	44	精準率 :0.52
	no	261	4646	召回率 :0.15
		hate	no	F1 score:0.23
		真實		

資料來源：本研究整理

4.3.2 3000_tweets_hate_goldlabel 資料集

本章節為使用 3000_tweets_hate_goldlabel 資料集，將該資料集如 4.1.2 所述，將原始資料集中馬來西亞、澳洲、美國三個資料集個別以 BERT 模型 Bi-LSTM 模型進行二分類任務訓練，再將三個資料集合併為一個資料集進行二分類任務訓練。

馬來西亞

表 22 為 3000_tweets_hate_goldlabel 資料集中馬來西亞資料集，以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終 loss 值。BERT 模型訓練了 3 個 epoch，Bi-LSTM 模型訓練了 20 個 epoch。由此可以看出，於測試集正確率的表現，使用 BERT 模型與使用 Bi-LSTM 模型並沒有太大的差異，使用 BERT 模型之測試集正確率為 0.78，使用 Bi-LSTM 模型之測試集正確率為 0.74。

而表 23 與表 24 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，在此我們可以發現，使用兩個模型的結果，除了測試集之正確

率類似外，於其他的評估標準來看，兩者的表現也極為類似，使用 BERT 模型的 F1 score 達到 0.51，而使用 Bi-LSTM 模型則僅有 0.52。

表 22 3000_tweets_hate_goldlabel 中馬來西亞資料集正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.78	0.79	0.52	0.52	45
Bi-LSTM	0.74	0.76	0.15	0.45	179

資料來源：本研究整理

1. BERT 模型

表 23 BERT、3000_tweets_hate_goldlabel 中馬來西亞資料集之混淆矩陣

預測				
	hate	27	19	精準率 :0.58
	no	27	138	召回率 :0.5
		hate	no	F1 score:0.51
		真實		

資料來源：本研究整理

2. Bi-LSTM 模型

表 24 Bi-LSTM、3000_tweets_hate_goldlabel 中馬來西亞資料集之混淆矩陣

預測				
	hate	30	24	精準率 :0.55
	no	29	128	召回率 :0.5
		hate	no	F1 score:0.52
		真實		

資料來源：本研究整理

澳洲

表 25 為 3000_tweets_hate_goldlabel 資料集中澳洲資料集，以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終

loss 值。BERT 模型訓練了 10 個 epoch，Bi-LSTM 模型訓練了 20 個 epoch。於測試集正確率的表現可以看出，與馬來西亞資料集的表現類似，兩個模型的測試集正確率沒有太大差異，使用 BERT 模型之測試集正確率為 0.84，而使用 Bi-LSTM 模型則為 0.82。

而表 26 與表 27 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，在此，則可以看出與馬來西亞資料集之差異，使用 BERT 模型的 F1 score 達到 0.81，而使用 Bi-LSTM 模型則僅有 0.77，兩者有些微的差距，而且與馬來西亞資料集使用兩個模型皆為 0.5 的結果，澳洲資料集有明顯較好的表現。

表 25 3000_tweets_hate_goldlabel 中澳洲資料集正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.84	0.88	0.63	0.63	86
Bi-LSTM	0.82	0.74	0.05	0.64	170

資料來源：本研究整理

1. BERT 模型

表 26 BERT、3000_tweets_hate_goldlabel 中澳洲資料集之混淆矩陣

預測				
	hate	71	10	精準率 :0.87
	no	22	108	召回率 :0.76
		hate	no	F1 score:0.81
		真實		

資料來源：本研究整理

2. Bi-LSTM 模型

表 27 Bi-LSTM、3000_tweets_hate_goldlabel 中澳洲資料集之混淆矩陣

預測				
	hate	64	7	精準率 :0.9
	no	29	111	召回率 :0.68
		hate	no	F1 score:0.77
		真實		

資料來源：本研究整理

美國

表 28 為 3000_tweets_hate_goldlabel 資料集中美國資料集，以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終 loss 值。BERT 模型訓練了 10 個 epoch，Bi-LSTM 模型訓練了 20 個 epoch。於測試集正確率的表現可以看出，與馬來西亞、澳洲資料集的表現不同，兩個模型的測試集正確率有些微的差異，使用 BERT 模型之測試集正確率為 0.80，而使用 Bi-LSTM 模型則為 0.74，使用 BERT 模型之測試集正確率表現較優。

而表 29 與表 30 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，在此可以看出，美國資料集為 3000_tweets_hate_goldlabel 中三個子資料集中表現最差的資料集，即使使用 BERT 模型之測試集 F1 score 表現較優於使用 Bi-LSTM 模型之 0.37，使用 BERT 模型之測試集 F1 score 也僅有 0.43，相較資料集中其他兩個子資料集的表現略低了許多。

表 28 3000_tweets_hate_goldlabel 中美國資料集正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.80	0.80	0.90	0.90	86
Bi-LSTM	0.74	0.85	0.11	0.40	193

資料來源：本研究整理

1. BERT 模型

表 29 BERT、3000_tweets_hate_goldlabel 中美國資料集之混淆矩陣

預測				
	hate	14	9	精準率 :0.60
	no	33	155	召回率 :0.29
		hate	no	F1 score:0.43
		真實		

資料來源：本研究整理

2. Bi-LSTM 模型

表 30 Bi-LSTM、3000_tweets_hate_goldlabel 中美國資料集之混淆矩陣

預測				
	hate	16	23	精準率 :0.41
	no	31	141	召回率 :0.34
		hate	no	F1 score:0.37
		真實		

資料來源：本研究整理

全部 3000

表 31 為將 3000_tweets_hate_goldlabel 資料集中的三個子資料集合併為一個資料集，以 8:2 分為訓練集與測試集，訓練集再以 8:2 分為訓練集與驗證集，使用 BERT 模型與 Bi-LSTM 模型進行訓練之測試集正確率、驗證集最終正確率、訓練集 loss 值與驗證集最終 loss 值。BERT 模型訓練了 10 個 epoch，Bi-LSTM 模型訓練了 40 個 epoch。於測試集正確率的表現可以看出，使用 BERT 模型與 Bi-LSTM 模型，並沒有太大的差異，前者之測試集正確率為 0.86，後者則為 0.85。

而表 32 與表 33 則各為兩個模型訓練後測試集之結果混淆矩陣與精準率、召回率和 F1 score，在此則可以看出，3000_tweets_hate_goldlabel 資料集，使用 BERT 模型的表現比使用 Bi-LSTM 模型較優，前者之 F1 score 為 0.68，後者則為 0.54，不過於精準率的部分，使用 Bi-LSTM 模型則達到 1，使用 BERT 模型則僅有 0.76。

表 31 3000_tweets_hate_goldlabel 資料集正確率、loss 值

	acc	eval_acc	loss	eval loss	訓練時間(秒)
BERT	0.86	0.84	0.71	0.71	206
Bi-LSTM	0.85	0.69	0.44	0.56	1074

資料來源：本研究整理

1. BERT 模型

表 32 BERT、3000_tweets_hate_goldlabel 資料集之混淆矩陣

預測				
	hate	85	27	精準率 :0.76
	no	54	443	召回率 :0.61
		hate	no	F1 score:0.68
		真實		

資料來源：本研究整理

2. Bi-LSTM 模型

表 33 Bi-LSTM、3000_tweets_hate_goldlabel 資料集之混淆矩陣

預測				
	hate	52	0	精準率 :1
	no	87	470	召回率 :0.37
		hate	no	F1 score:0.54
		真實		

資料來源：本研究整理

4.3.3 綜合結果

表34為HatebaseTwitter資料集之綜合結果，而表35為3000_tweets_hate_goldlabel 資料集之綜合結果，由此二表可知，使用BERT模型不論是使用HatebaseTwitter資料集或3000_tweets_hate_goldlabel 資料集，表現皆較優於使用Bi-LSTM模型，不過兩者有些差異，使用HatebaseTwitter資料集時BERT模型與Bi-LSTM模型的差異較為明顯，而使用3000_tweets_hate_goldlabel 資料集時，BERT模型表現雖還是較優，但差異卻較小，我們認為，這個差異的原因在於仇恨言論佔整個資料集的比例，這兩個資料集雖然皆為極度不平衡的資料集，但是仇恨言論所佔的比例還是有差別，HatebaseTwitter 資料集中仇恨言論佔了約 5%，3000_tweets_hate_goldlabel 資料集中仇恨言論佔了約 22%，後者比例還是較高的緣故，Bi-LSTM模型對於處理這種極度不平衡的資料集的能力，相較BERT模型還是較差的。

表 34 HatebaseTwitter 資料集綜合結果

資料集		BERT	Bi-LSTM
2 分類	正確率	0.93	0.93
	F1-score	0.44	0.23
3 分類	正確率	0.90	0.76
	F1-score	0.48	0

資料來源：本研究整理

表 35 3000_tweets_hate_goldlabel 資料集綜合結果

資料集		BERT	Bi-LSTM
馬來西亞	正確率	0.78	0.74
	F1-score	0.51	0.52
澳洲	正確率	0.84	0.82
	F1-score	0.81	0.77
美國	正確率	0.80	0.74
	F1-score	0.43	0.37
全部 3000	正確率	0.86	0.85
	F1-score	0.68	0.54

資料來源：本研究整理

3000_tweets_hate_goldlabel 資料集將三個子資料集合併的結果 F1-score 達到 0.68，而其個別子資料集澳洲的部分甚至達到 0.81，相較於使用 HatebaseTwitter 資料集的 F1-score 三分類僅有 0.44，而二分類則僅有 0.48，明顯較優，儘管如此，由於 3000_tweets_hate_goldlabel 資料集並非為公開之資料集，我們無法與其他研究之結果進行比較。

然而，HatebaseTwitter 資料集在過去則有其他研究使用進行仇恨言論偵測，該資料集原始論文中，(Davidson et al., 2017) 使用了邏輯斯迴歸模型(Logistic regression)跟 L2 正規化(regularization) 進行三分類預測任務，針對仇恨言論的 F1-score 達到 0.51，而另一個研究，(Zhang & Luo, 2018)的研究同樣使用了 HatebaseTwitter 資料集，將資料及分為仇恨言論(hate)與非仇恨言論(non-hate)兩類，使用預訓練完成的 word2vec 模型，預訓練語料為 Google 新聞語料，預訓練模型為 Skip-gram 模型，將仇恨言論語料轉換為詞向量作為該研究多種模型之輸入，而其研究之結果針對仇恨言論的 F1-score 為下，使用了 SVM 為 0.23、CNN+sCNN 為 0.3、CNN+GRU 為 0.29 與其他幾種方法，針對仇恨言論的 F1-score 皆不超過 0.3，而本研究二分類任務使用 BERT 模型則達到 0.44。

我們也發現，仇恨言論佔資料集的比例將會影響到神經網路訓練的結果，下表為我們使用的各個資料集仇恨言論所佔比例。

表 36 仇恨言論佔資料集比例

資料集	仇恨言論所佔比例
HatebaseTwitter(三分類)	5.7%
HatebaseTwitter(二分類)	5.7%
馬來西亞	23.9%
澳洲	45%
美國	19.4%
全部 3000	25.6%

資料來源：本研究整理

由表 36 與表 35、34，我們可以看出，仇恨言論佔整個資料集的比例越平衡，將得到越好的實驗結果，澳洲資料集擁有全部六個資料集最平衡的比例 45%，因此也獲得了最好的 F1-score 結果 0.81，因此，我們認為對於 Hatebase-Twitter 資料集的研究針對仇恨言論的 F1-score 表現皆偏低的原因，還是在於仇恨言論所佔的比例太低的緣故，僅有 5%的仇恨言論，將導致神經網路或是其他簡單的分類器無法將仇恨言論正確的預測出來。

第五章 結論

5.1 結論

本研究以兩個不同的資料集，內含仇恨言論的標記，進行仇恨言論的偵測分類，將資料集原始之資料分配以及將資料集進行調整後，分別為 HatebaseTwitter 資料集原始三分類、以及調整後之二分類，以及 3000_tweets_hate_goldlabel 資料集之三個個別子資料集，以及將三個子資料集合併，共六種資料集，使用兩個不同的深度學習模型，BERT 模型與 Bi-LSTM 模型，進行訓練，綜合兩個模型結果進行評估。

而最終，兩個模型給予相同的資料集時，即使於大部分的結果中，以正確率來看並沒有顯著的差異，但當從我們使用的其他評估標準如精準率、召回率和 F1 score 來看，使用 BERT 模型的結果表現較優於使用 Bi-LSTM 模型，Hatebase-Twitter 資料集使用 BERT 模型進行三分類任務之 F1-score 達到 0.44，而 3000_ tweets_hate_goldlabel 資料集整體的 F1-score 達到 0.68，其子資料集澳洲的部分甚至達到 0.81。我們也發現對於仇恨言論所佔的比例越少，使用 BERT 模型的結果與使用 Bi-LSTM 模型的差距會更明顯。我們也認為，仇恨言論所佔比例將會影響神經網路訓練的成果，這也是為什麼 HatebaseTwitter 資料集不論是本研究或其他研究的結果都不甚理想，該資料集中仇恨言論所佔的比例過低，僅佔其中 5%，這將導致神經網路很難正確預測出仇恨言論。

5.2 研究限制

仇恨言論資料集取得不易，過去許多仇恨言論偵測的相關研究中，並沒有公開其資料集，或是礙於社群網站如 twitter 之相關規範，公布之資料集並不包含實際 twitter 推文。若是為自行收集仇恨言論資料集，在判斷是否仇恨言論時，部分言論仍會存在爭議。

且由於仇恨言論於真實的網路世界中，所佔的比例還是相對很少，這樣極端不平衡的資料集，會導致深度學習模型的訓練成果有限。如本研究使用之兩個資料集，就因為仇恨言論所佔之比例不同，造成不同之結果。

5.3 未來研究方向

本研究只有使用到一個公開資料集與一個非公開資料及進行比較，且皆為 Twitter 之言論，未來研究之可蒐集其他不同社群媒體之仇恨言論資料集，進行訓練與比較。而近年來，許多的語言模型不斷的被改良、提出，如：ALBERT、RoBERTa、XLNet 等，未來研究可以將這些模型與 BERT 語言模型進行比較。

參考文獻

[1] Agarwal, S., & Sureka, A. (2015). *Using knn and svm based one-class classifier for detecting online radicalization on twitter.* Paper presented at the International Conference on Distributed Computing and Internet Technology.

[2] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). *Deep learning for hate speech detection in tweets.* Paper presented at the Proceedings of the 26th International Conference on World Wide Web Companion.

[3] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137-1155.

[4] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., . . . Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change, 95*, 96-108.

[5] Burnap, P., & Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

[6] Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 7*(2), 223-242.

[7] Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies, 65*(1), 57-70.

[8] Chen, Y. (2011). Detecting offensive language in social medias for protection of adolescent online safety.

[9] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety.* Paper presented at the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing.

[10] Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). *Improving cyberbullying detection with user context.* Paper presented at the European Conference on Information Retrieval.

[11] Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *arXiv preprint arXiv:1905.12516*.

[12] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech*

detection and the problem of offensive language. Paper presented at the Eleventh international aaai conference on web and social media.

[13] Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook.

[14] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[15] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(3), 18.

[16] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). *Hate speech detection with comment embeddings.* Paper presented at the Proceedings of the 24th international conference on world wide web.

[17] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 85.

[18] Gambäck, B., & Sikdar, U. K. (2017). *Using convolutional neural networks to classify hate-speech.* Paper presented at the Proceedings of the first workshop on abusive language online.

[19] Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215-230.

[20] Guermazi, R., Hammami, M., & Hamadou, A. B. (2007). *Using a semi-automatic keyword dictionary for improving violent Web site filtering.* Paper presented at the 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System.

[21] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

[22] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

[23] Ji, S., Yun, H., Yanardag, P., Matsushima, S., & Vishwanathan, S. (2015). Wordrank: Learning word embeddings via robust ranking. *arXiv preprint arXiv:1506.02761*.

[24] Kwok, I., & Wang, Y. (2013). *Locate the hate: Detecting tweets against blacks.* Paper presented at the Twenty-seventh AAAI conference on artificial intelligence.

[25] Lomas, N. (2017). Facebook, google, twitter commit to hate speech action in germany. *Last accessed: July*.

[26] MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one, 14*(8).

[27] McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs, 77*(2), 257-280.

[28] Mehdad, Y., & Tetreault, J. (2016). *Do characters abuse more than words?* Paper presented at the Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

[29] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[30] Mishna, F., Cook, C., Saini, M., Wu, M.-J., & MacFadden, R. (2011). Interventions to prevent and reduce cyber abuse of youth: A systematic review. *Research on Social Work Practice, 21*(1), 5-14.

[31] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). *Abusive language detection in online user content.* Paper presented at the Proceedings of the 25th international conference on world wide web.

[32] Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution, 3*, 1277-1279.

[33] Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

[34] Patchin, J. W., & Hinduja, S. (2012). *Cyberbullying prevention and response: Expert perspectives*: Routledge.

[35] Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). *Deeper attention to abusive user content moderation.* Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[36] Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

[37] Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). *Offensive language detection using multi-level classification.* Paper presented at the Canadian Conference on Artificial Intelligence.

[38] Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using*

natural language processing. Paper presented at the Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.

[39] Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). *Analyzing the targets of hate in online social media.* Paper presented at the Tenth International AAAI Conference on Web and Social Media.

[40] Sood, S., Antin, J., & Churchill, E. (2012). *Profanity use in online communities.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[41] Spertus, E. (1997). *Smokey: Automatic recognition of hostile messages.* Paper presented at the Aaai/iaai.

[42] Thompson, N. (2016). *Anti-discriminatory practice: Equality, diversity and social justice*: Macmillan International Higher Education.

[43] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., . . . Hoste, V. (2015). *Detection and fine-grained classification of cyberbullying events.* Paper presented at the Proceedings of the international conference recent advances in natural language processing.

[44] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is all you need.* Paper presented at the Advances in neural information processing systems.

[45] Wang, X., Gerber, M. S., & Brown, D. E. (2012). *Automatic crime prediction using events extracted from twitter posts.* Paper presented at the International conference on social computing, behavioral-cultural modeling, and prediction.

[46] Warner, W., & Hirschberg, J. (2012). *Detecting hate speech on the world wide web.* Paper presented at the Proceedings of the second workshop on language in social media.

[47] Waseem, Z. (2016). *Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter.* Paper presented at the Proceedings of the first workshop on NLP and computational social science.

[48] Waseem, Z., & Hovy, D. (2016). *Hateful symbols or hateful people? predictive features for hate speech detection on twitter.* Paper presented at the Proceedings of the NAACL student research workshop.

[49] Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus.* Paper presented at the Proceedings of the 21st ACM international conference on Information and

knowledge management.

[50] Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). *Learning from bullying traces in social media.* Paper presented at the Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies.

[51] Yuan, S., Wu, X., & Xiang, Y. (2016). *A Two Phase Deep Learning Model for Identifying Discrimination from Tweets.* Paper presented at the EDBT.

[52] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. *arXiv preprint arXiv:1902.09666*.

[53] Zhang, Z., & Luo, L. (2018). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*(Preprint), 1-21.

[54] Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., & Caragea, C. (2016). *Content-Driven Detection of Cyberbullying on the Instagram Social Network.* Paper presented at the IJCAI.

[55] 李洋, & 董红斌. （2018）. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析. *计算机应用, 38*（11）, 3075-3080.