The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 92-101
©The Association for Computational Linguistics and Chinese Language Processing

結合 LDA 與 SVM 之社群使用者立場檢測

Stance Detection of Social Network Users by combining Latent Dirichlet Allocation and Support Vector Machine

翁翊桓　I-Huan Weng
國立臺北科技大學資訊工程學系
Department of Computer Science and Information Engineering
National Taipei University of Technology
t106598025@ntut.org.tw

王正豪　Jenq-Haur Wang
國立臺北科技大學資訊工程學系
Department of Computer Science and Information Engineering
National Taipei University of Technology
jhwang@csie.ntut.edu.tw

摘要

傳統的立場分析常常使用問卷調查、電話訪查等來得知不同的主題下每個人的觀點。但由於傳統的統計方法，採用抽樣的方式，容易因為樣本數的不足，導致效果較差。現有的方法包括以情緒字典、以卷積式類神經網路(CNN)、遞歸式類神經網路(RNN)等，但是因為深度類神經網路需要較多資料集才能提升效果。而文本的特徵則採用 N-Gram 或是 TF-IDF 方法，但這樣無法真正了解文本的語意。本論文提出利用 Word2Vec 字詞表示模型，來取得字詞的向量，並結合 LDA 方法來取得文本的特徵。在立場檢測方面，我們以 SVM 作為分類器，以兩階段方法分辨人們是否中立與否的主觀性問題，並預測使用者的立場。

本論文以 SemEval-2016 的立場偵測任務，作為實驗的資料來源，並使用多種方法(F-Measure, Accuracy, Precision, Recall)來評估效果，相較於 SemEval-2016 的基線或其他隊伍分數，平均而言，本論文所提的方法皆獲得較好的結果 (F-Measure : 83.36%)。

Abstract

In traditional stance analysis, questionnaire survey or telephone survey are often used to know the opinions of each person under different topics. However, due to the traditional statistical methods, the sample size is too small to get good result. Existing methods are usually based on sentiment lexicon, Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN). And the text features are based on N-Gram or TF-IDF, which do not help to understand the semantics of the text. This research proposes to use Word2Vec for word embedding and combine the LDA to obtain the text feature. For stance detection, we use Support Vector Machine (SVM) to train the classifier to detect the subjectivity of texts, and to predict user stances.

In the experiment, we used the data from SemEval-2016 Stance Detection Task, and use a variety of evaluation methods (F-Measure, Accuracy, Precision, Recall) to evaluate performance. Compared with SemEval-2016 official baseline and other teams scores, our proposed method can get better result on average (F-Measure : 83.36%).

關鍵詞：立場檢測、機器學習、社群網路分析、Word2vec、隱含狄利克雷分布

Keywords: Stance Detection, Machine learning , Social network analysis , Word2vec

一、緒論

人們在生活中針對的目標不同時，人們的立場也會變得不一樣，通常人們有所衝突時大多數都是因為立場不同，像是電影評論、產品意見、總統大選等有關的問題。這些問題通常都會被人所收集來進行探討與分析。早期的方法通常是以電訪或紙本問卷來進行抽樣的調查，所以容易因為人力與樣本數量的關係，進而影響到預測的結果。

近年來，網際網路的普及，造就了許多的社群網路平台像是：Twitter、Facebook，而根據 Statista 的數據統計[1]指出：全球社群媒體的使用者在 2019 年估計有 27.7 億人，而光是在台灣，社群網站的使用者數量就佔了總人口 89%，能得知社群網站對於人們來說成為了生活中不可或缺的存在。使用者平常會在這些平台發表自己對不同主題的想法或是意見，因此每天都會有許多訊息存在於社群上。若以傳統的方法來針對這些大量的資訊來對使用者與貼文進行分析，除了需要配置許多人力與成本花費，還需要長時間才能有所結果，而且新的訊息還會隨著時間大量增加，因此透過機器學習的技術與系統自動

化來處理數據的分析將是未來的趨勢。

由於現有文本處理通常都只以提用詞移除(Stopword removal)作前置處理，因此本論文另外再使用詞型還原(Lemmatization)與詞幹提取(Stemming)的方法來評估對於分類器的影響。在立場檢測上，本論文提出以透過 LDA 的方法，來產生主題特徵並與經由 Word2Vec 所轉換的特徵向量作結合，來進行立場的檢測。

最後，根據本論文之實驗，在使用 LDA 結合特徵向量時，確實能使各個目標主題的準確率提升，最高的 F-Measure 為 84.23%，平均 F-Measure 為 76.88%皆高於 SemEval-2016 上的 Baselines 與其他隊伍，因此可以驗證所提出方法是有效的。

二、相關研究

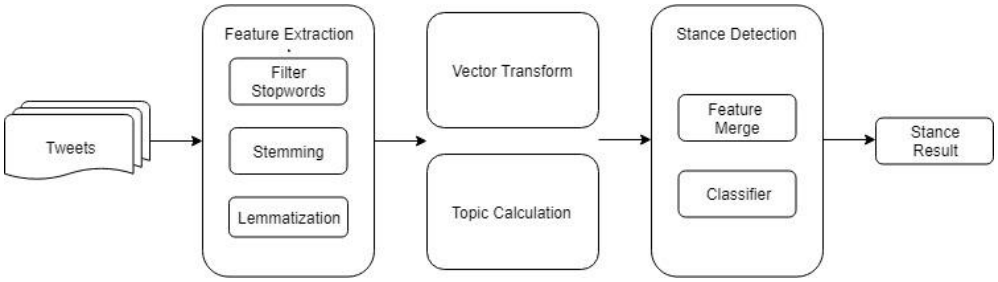
Jannati 等人[2]透過收集部落格的文章來檢測對於政治人物的立場，並以情緒辭典作分析；Tumasjan 等人[3]對 Twitter 上的貼文進行分析並依其結果來預測 2009 年的德國總理大選，來證實社群平台上的貼文確實可以反映出人們的立場；Sasaki[4]等人則是對 Twitter 的貼文添加額外的提示標籤，實驗發現透過添加的額外事件特徵可以提升立場檢測的準確度。 Tutek 等人[5]透過 SVM 對文本作 N-Gram 來進行立場檢測；Bøhler 等人[6]使用 GloVe 詞向量模型，比較了 Naive Bayes, SVM 這兩種分類演算法的效果，發現結合 GloVe 詞向量的特徵能提高檢測效果；[7, 8]使用卷積神經網路(CNN)來進行立場檢測，其中使用 Semeval-2016 的資料集在 F-measure 有 67.33%；Zarrella 等人[9]透過遞歸神經網路(RNN)並使用預訓練的特徵來進行立場檢測，使用 Semeval-2016 的資料集在 F-measure 有 67.8%。根據 Igarashi 等人[10]的觀察發現，深度學習方面效果沒有一般的分類好；而 Mohammad 等人[11]也發現在立場檢測上，SVM 的整體平均高於其他模型。基於資料集的數量與前者的觀察，我們採用 SVM 作為主要的方法進行立場檢測，並在實驗中與 Naive Bayes, LSTM 等分類器作比較。

在相關的論文中，大部分的平台都是以 Facebook 或是 Twitter 平台為主作分析，且主題大多都圍繞在政治上。而目前常用的方法是使用針對文本中目標的情緒極性，這樣方法有些缺點，像在單一目標或跨領域時效果通常不佳，為了提升立場的主題多樣性和準確率，本論文將利用 LDA 主題模型的特性，結合 LDA 立場模型與 Word embedding，期

望能達到更良好的結果。

三、研究方法

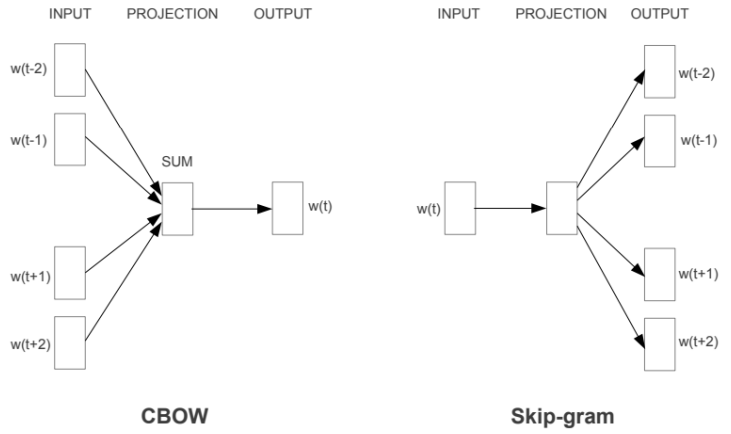
本研究所提出之方法主要可以分為四大部分，一開始會接收文本的輸入，並透過 Feature Extraction 來對輸入的文本進行 Filter Stopwords、Stemming、Lemmatization 來取得文本的特徵。之後分為兩個階段進行，第一階段會透過 Topic Calculation，以 LDA 主題模型來取得各個文本的主題特徵，並對各個文本做主題分佈的標記；第二階段則是透過 WordVec[13]字詞向量空間模型，將各個文本的字詞轉成向量表示。最後合併第一階段與第二階段的結果，透過監督式學習法來訓練分類器並取得立場結果。



圖一、系統架構圖

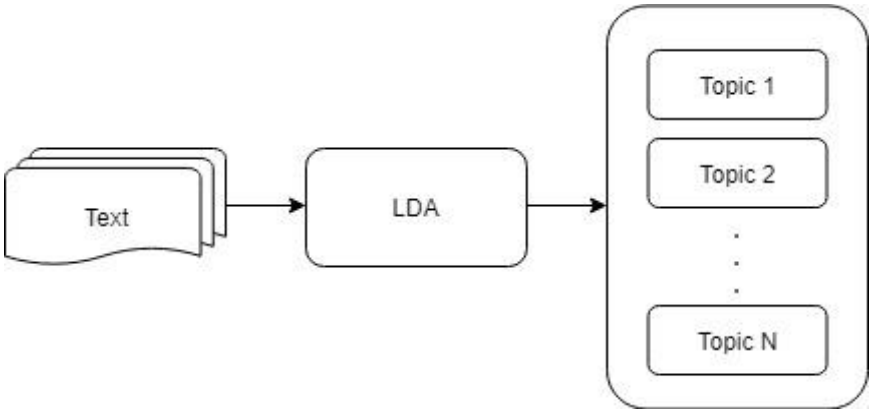
一個文本是由許多的文字所組成的，而文本中有許多字其實是不必要的，從文本中找到有意義或是重要的字，即可獲得較好的數據來進行分析。本研究從文本中移除 Stopwords，接著以詞幹提取(Stemming)、詞性還原(Lemmatization)等方法，來取得文本的完整語義。本研究使用 NLTK 中的斷詞 Stopwords 清單，過濾掉屬於停用詞的文字來提供後續的步驟作使用。詞幹提取是透過抽取字詞的詞綴來獲得詞根的方式，目的是讓變化的字詞簡化，使得文本分析能獲取到較好的字義來使用，本研究使用被廣為接受的 Porter Stemmer 來進行詞幹提取。詞型還原能把一個任何型別的字詞還原為原型，目的是能讓將字詞簡化為最初的字詞原形，使得文本分析能獲得字詞的完整語義，用在更為精確的自然語言處理上的文本分析與表達。將文本的內容透過機器學習進行分類前，必

須先將文本的內容轉換成向量，作為訓練分類器前的輸入。為了表示字詞的語義關係，本研究使用預先訓練好的 Word2Vec 進行文本的向量轉換，以便處理後續的實驗與步驟。Word2Vec[14]是由 Google 的 Tomas Mikolov 等人於 2013 年所提出的一種 Word Embedding model，在 Word2Vec 中，透過神經網路的方法，將文本中的文字與文字的關係轉化成具有語義關係和語法結構的向量形式。Word2Vec 中有兩個模型，如圖二，一個為連續型詞袋模型(CBOW)以及跳躍式模型(Skip-gram)兩種。CBOW 是透過輸入的上下文來預測字詞，而 Skip-gram 是透過輸入的字詞來預測上下文。



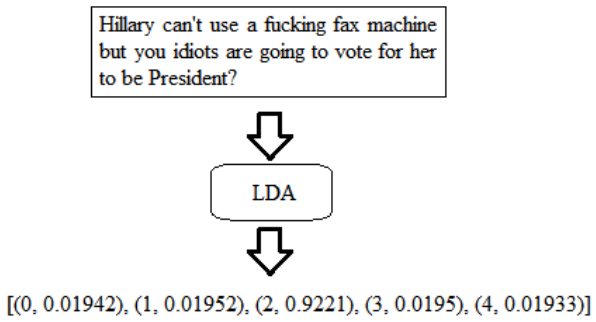
圖二、CBOW 以及 Skip-gram 之 Word Embedding 模型示意圖

在訓練字詞模型的時間上，Skip-gram 相較於 CBOW 的訓練時間會較久，但是在語意分析上，Skip-gram 比 CBOW 還來得好，本研究所使用的 Google 訓練好的模型，是基於 Skip-gram 上作訓練的模型。本研究使用經過前處理的文本，並利用主題模型，對每篇文本作標記，來取得主題特徵。在這一篇章節中，將會說明如何取得，並使之當作本研究特徵。隱含狄利克雷分布(Latent Dirichlet Allocation)簡稱 LDA，是由 Blei[12]等人在 2003 年所提出的主題詞袋模型。利用不同的機率的潛在主題，來描述每篇文章，而每一個主題是由分散的主題字詞所形成的。LDA 主題模型也是一個生成模型，他將每篇文本的主題按照不同的機率來表現每篇文章。LDA 的生成步驟大致上為圖三所示，輸入文本，之後經由 LDA 輸出主題數 N 的各個主題。



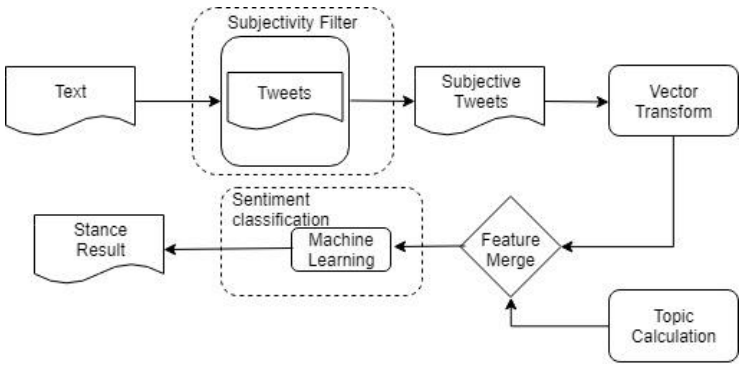
圖三、LDA 生成步驟

主題特徵的擷取步驟為圖四所示，輸入一篇文本，經由 LDA 取得該文本在各個主題中的分布概率。本研究採取文本的主題分布概率來作為主題特徵，並與所產生的字詞向量特徵進行結合，來進行後續的實驗。



圖四、LDA 文本主題分布概率示意圖

為了找出文本作者在貼文中所隱含的立場，我們使用機器學習的方式來針對貼文的內容作立場的分類，在這邊透過兩階段的方法，透過找出貼文者的主觀性立場與情感極性，來預測使用者的立場。



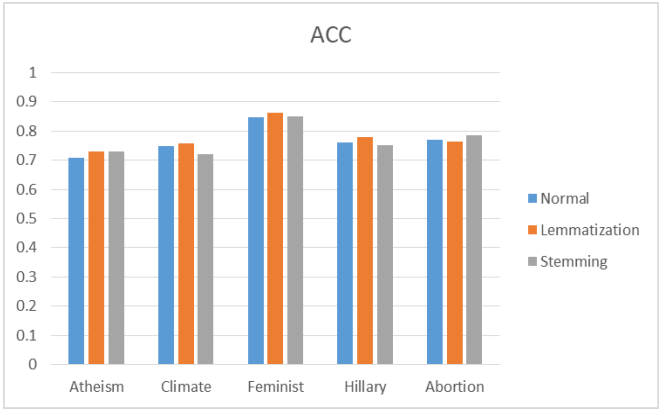
圖五、Stance Detection 架構圖

在立場判斷上，推文的人對於事情的主觀與客觀極為重要，具有中立立場的推文具有非主觀性的看法，而在這邊具有支持與反對的訊息則會有非中立情緒的要素，在這階段如圖五所示，因為有可能會對分類模型的訓練產生影響，所以我們在這對文本作過濾，將文本區分成含中立的文本和不含中立的文本。本研究使用 Scikit-learn 工具來進行 SVM 極感極性分類，在進行分類前，必須將訓練資料和測試資料轉為所需的格式。如圖五，我們所處理完的文本經由向量轉換與主題特徵擷取的結合，透過 SVM 來進行情感極性的判斷。

四、實驗方法

本研究的資料集是使用 SemEval-2016[11]在 Task6 中所提供的測試與訓練資料，共 4063 筆。資料裡共有五個主題分別為：Atheism、Climate Change、Feminist Movement、Hillary Clinton、Legal. Abortion 等，內容為推特使用者針對各主題表達自身的評論或想法。

在機器學習上，文本的處理極為重要，在轉換為向量以前，如果一篇文本的雜訊過多，那麼就會影響到後續的輸入。因此在本部分的實驗，我們將以不同的文字處理方法來處理文本，並比較他們對於 SVM 分類的影響。下圖 4.2 分別有三種處理方法，Normal 為停用詞過濾的基本文本處理；Lemmatization 為詞型還原；Stemming 為詞幹提取。

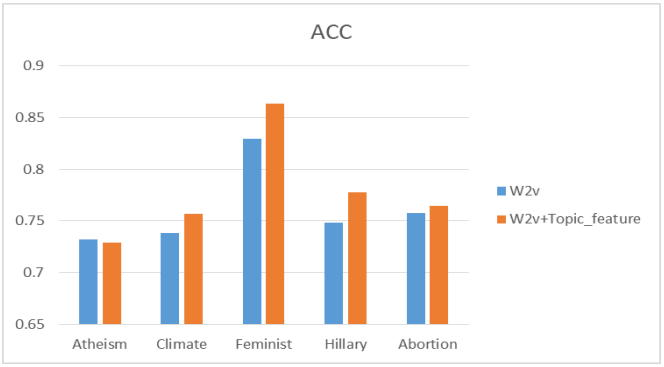


圖六、不同文本處理之分類準確率

由圖六得知，橫軸為各個目標主題，縱軸為準確率，經由詞型還原的資料，在分類的準

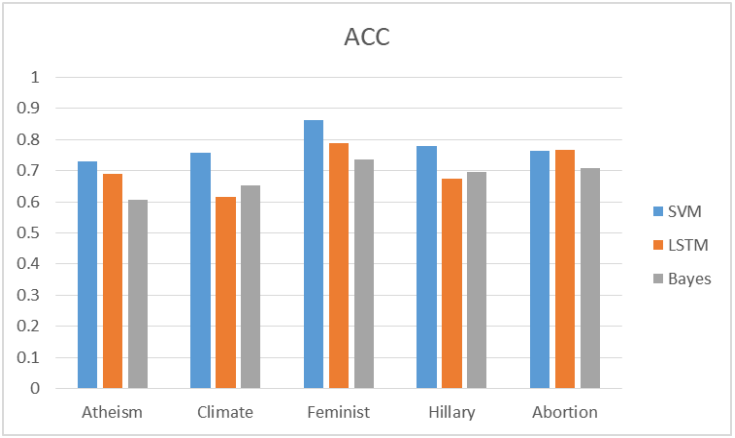
確率上大多有所提升，而一般的處理與詞幹提取的資料多比經由詞型還原的資料的準確率低。

從前面實驗我們知道詞型還原比其他文本處理的方法好，因此我們將使用詞型還原的文本處理方法來做主題特徵的實驗，本實驗將把主題特徵與 Word Embedding 的向量結合，進而與未添加主題特徵的一般字詞向量特徵做比較。



圖七、主題特徵實驗比較之準確率

由圖七得知，橫軸為各個目標主題，縱軸為準確率，我們可以發現 Feminist Movement 與 Hilary Clinton 在添加主題特徵後有明顯提升準確率，在 Legal. Abortion、Climate Change 略為增加；在 Atheism 上則沒明顯增加，原因可能在 Feminist 與 Hillary 的議題通常帶有高主觀性與高情緒極性的用字，而 Atheism 的議題則常出現一些情緒字眼較不明顯的用字。為了驗證分類模型的效果，所以我們以原先的資料集進行測試，並與樸素貝氏分類(Bayes)、長短期記憶神經網路分類模型(LSTM)這幾種方法來進行比較。



圖八、不同分類器之分類準確率

由圖八得知，橫軸為各個目標主題，縱軸為準確率，可以知道三種分類器中，SVM 最為優秀，而 LSTM 與 Bayes 分類器則較差。根據 Igarashi 等人[11]的實驗表示，深度學習的方法不見得會比較好。

五、結論

本論文提出使用 Word2Vec 字詞向量，結合主題特徵來進行立場檢測。在文本處理方面，使用三種文本處理方式，當中的 Lemmatization(詞型還原)於實驗中得證，在文本分析上的準確率較優。在主題特徵的方面，透過隱含狄利克雷分布(LDA)的方法來算出文本的主題分布能有效提升分類效果。在立場檢測方面，本論文透過支援向量機來訓練立場分類器，並且使用兩階段的方法來解決主觀性的問題，經實驗驗證最佳的 F-Measure 為主題目標女權運動的 83.36%。

參考文獻

[1] Statista, Number of social media users worldwide from 2010 to 2020,https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Viewed on 2019/07/02)

[2] Jannati, R., Mahendra, R., Wardhana, C. W., & Adriani, M. (2018, November). Stance Classification Towards Political Figures on Blog Writing. In 2018 International Conference on Asian Language Processing (IALP) (pp. 96-101).

[3] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth international AAAI conference on weblogs and social media.

[4] Sasaki, A., Mizuno, J., Okazaki, N., & Inui, K. (2016, October). Stance classification by recognizing related events about targets. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 582-587). IEEE.

[5] Tutek, M., Sekulic, I., Gombar, P., Paljak, I., Culinovic, F., Boltuzic, F., ... & Šnajder, J.

(2016, June). Takelab at semeval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)(pp. 464-468).

[6] Bøhler, H., Asla, P., Marsi, E., & Sætre, R. (2016, June). Idi@ ntnu at semeval-2016 task 6: Detecting stance in tweets using shallow features and glove vectors for word representation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 445-450).

[7] Wei, W., Zhang, X., Liu, X., Chen, W., & Wang, T. (2016, June). pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 384-388).

[8] Vijayaraghavan, P., Sysoev, I., Vosoughi, S., & Roy, D. (2016, June). DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 413-419).

[9] Zarrella, G. and A.J.a.p.a. Marsh, Mitre at semeval-2016 task 6: Transfer learning for stance detection. 2016.

[10] Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., & Inui, K. (2016, June). Tohoku at SemEval-2016 task 6: feature-based model versus convolutional neural network for stance detection. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 401-407).

[11] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 31-41).

[12] Blei, D.M., A.Y. Ng, and M.I.J.J.o.m.L.r. Jordan, Latent dirichlet allocation. 2003. 3(Jan): p. 993-1022.

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

[14] Google-News-dataset,https://code.google.com/archive/p/word2vec/ (Viewed on 2019/07/02)