以PTT八卦版文字分析探討 黑暗榮耀引爆之霸凌議題

第12組 期中報告

組長: N104020027 林欣穎

組員: N104020005 李姿儀

N104020024 林品均

N104020023 林義行

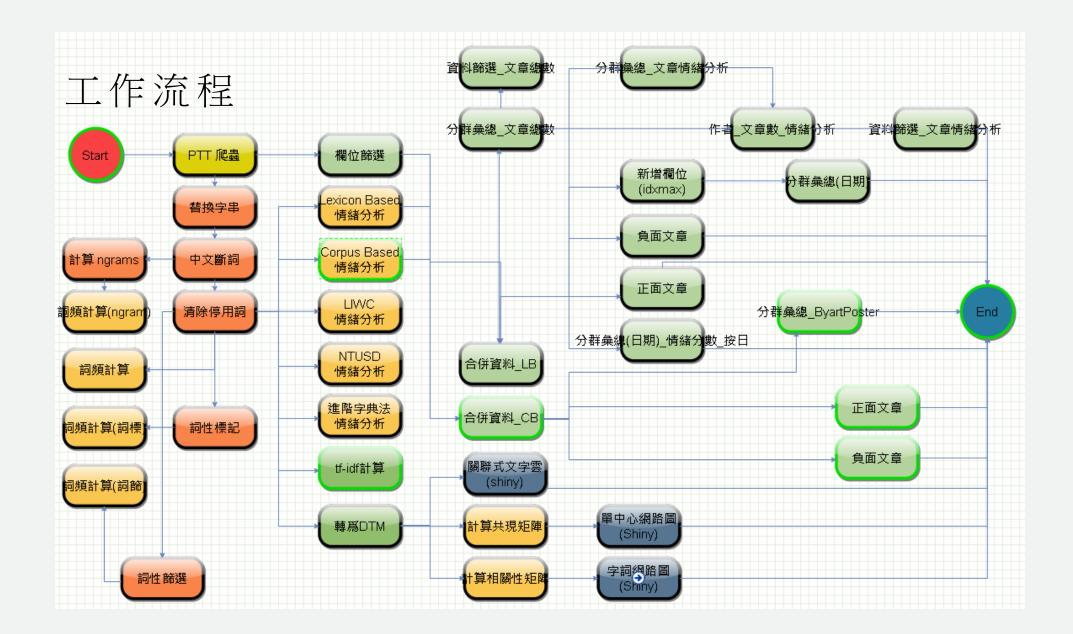
AGENDA

- 議題簡介
- 工作流程
- 資料清理
- 情緒分析
- 文辭及文件分析
- 視覺化圖表分析
- 小結

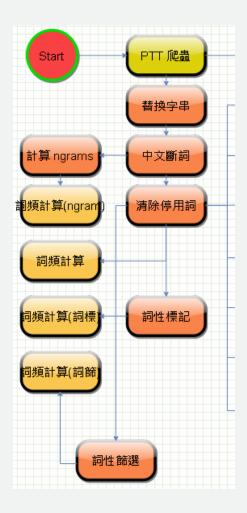
議題簡介

從校園霸凌到網路霸凌,大大小小的霸凌事件不斷在現實中發生,霸凌不只對受害者當下有深刻身心靈的傷害,對於社會本身也是一種重大的挑戰,例如近日非常火紅 Netflex 原創韓劇-黑暗榮耀,就是講述著校園中霸凌的故事。本組期望透過文字分析,獲得一些不論對他人或自我有所啟發的結果。





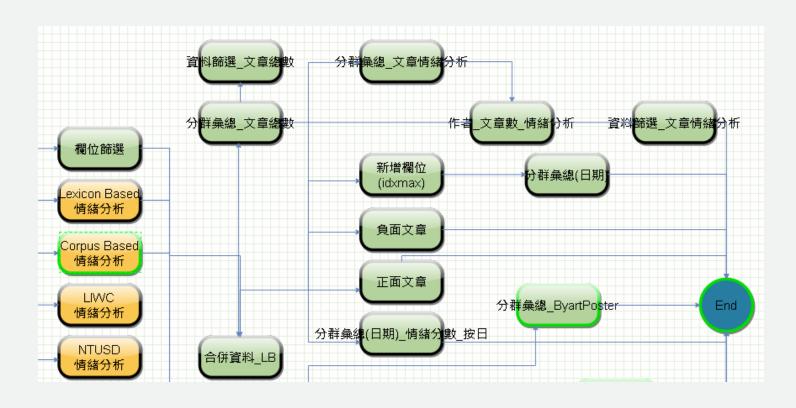
資料清理



替換字串 中文斷詞 清除停用字

- 透過替換字串清理文章內的雜訊,例如 Html 符號、網址以 及從何種裝置發送的字詞
- ▶ 進行中文斷詞,為霸凌、學校等字詞設定權重
- ➤ 斷詞後的結果透過 ngram 計算詞頻,用來預測後面通常接哪 些詞彙,並依此放入自訂斷詞裡,進行詞彙定義的權重,或 是放入自定義的停用詞裡

情緒分析-LEXICON BASED

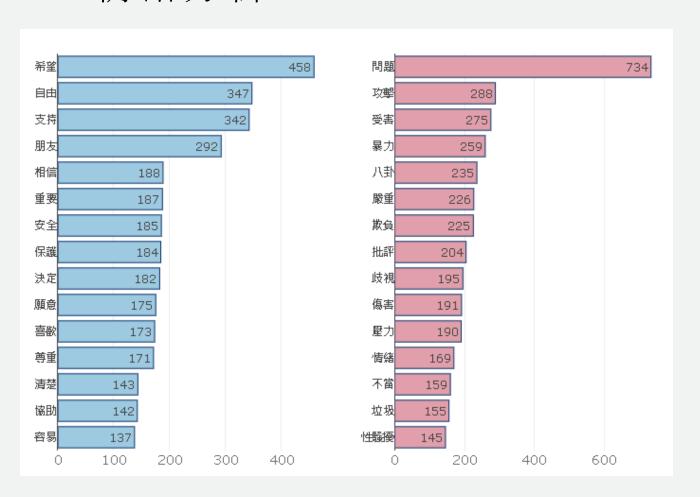


Lecicon Based 情緒分析

與原始文章 合併 依照作者計算 發文數量 挑選發文總數超過 10 篇的作者

計算作者的情 緒分數總和

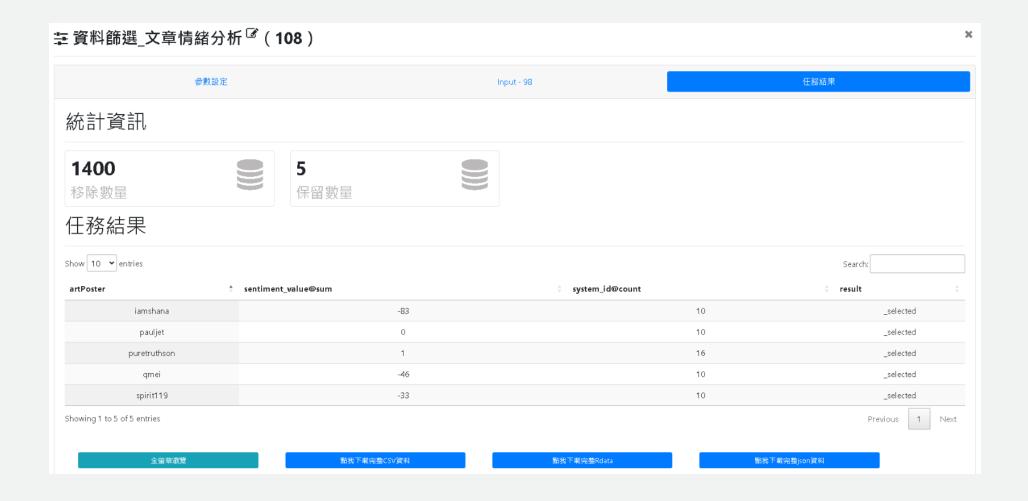
情緒分析-LEXICON BASED



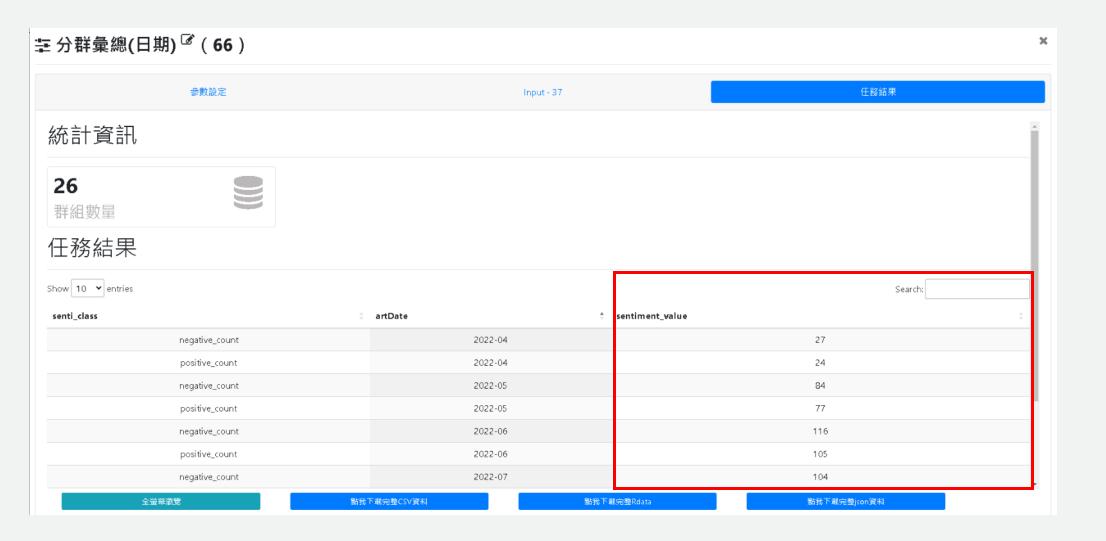


將個別發文者的發文 總數量與情緒總分數 進行合併

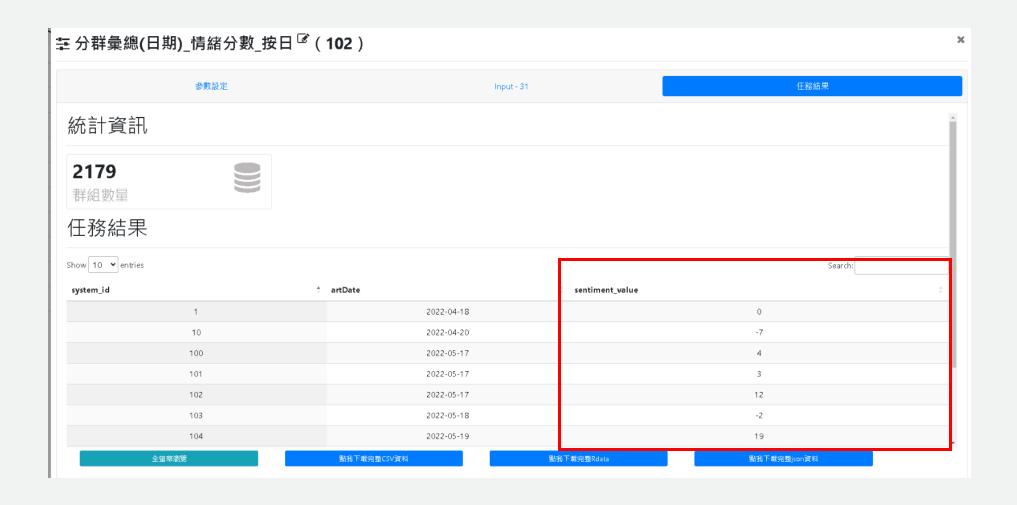
挑選出發文總數大 於10篇的發文者



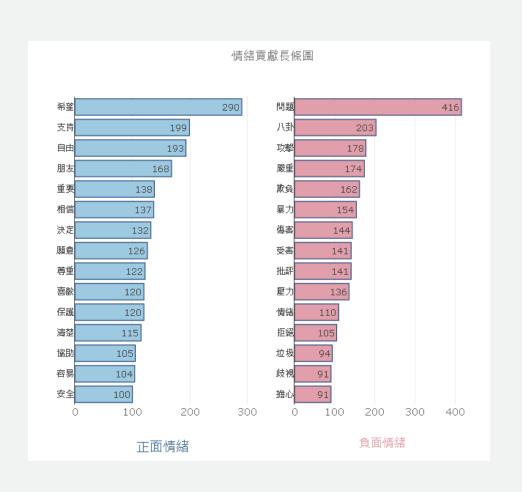
新增欄位用來區 依月份分別計算 分文章的正負面 正負文章總數

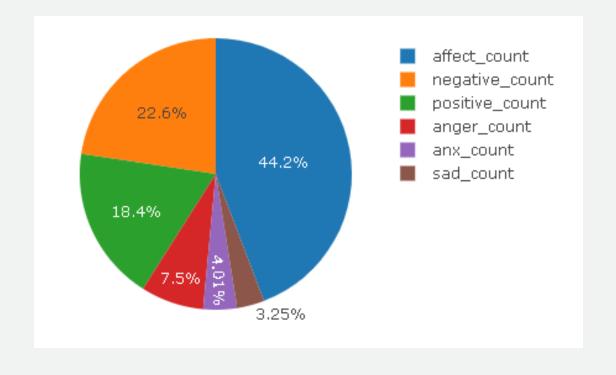


依每日分別計算 情緒分數

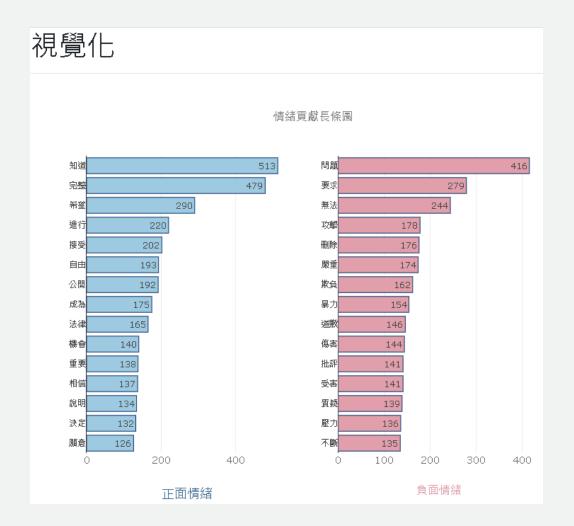


LIWC情緒分析





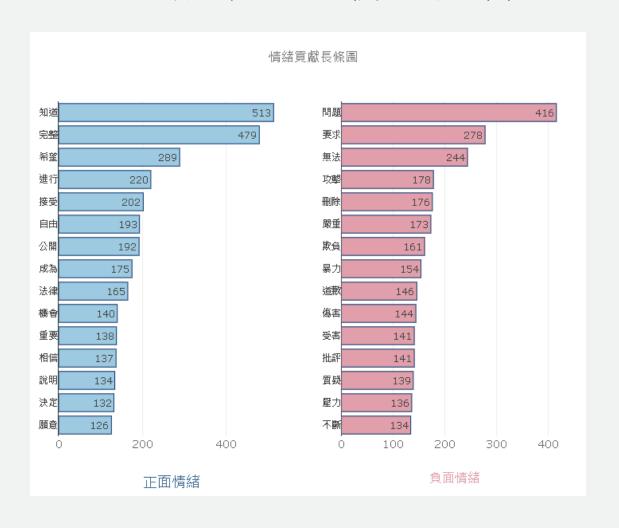
NTUSD情緒分析





- 「希望」一詞也在前三名字詞中出現。
- · 「問題」一詞佔據負面 情緒詞彙中的榜首。

進階字典法情緒分析

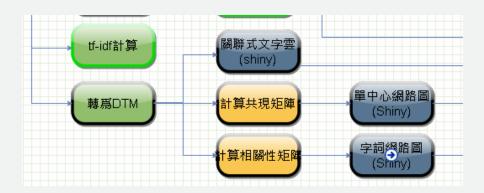


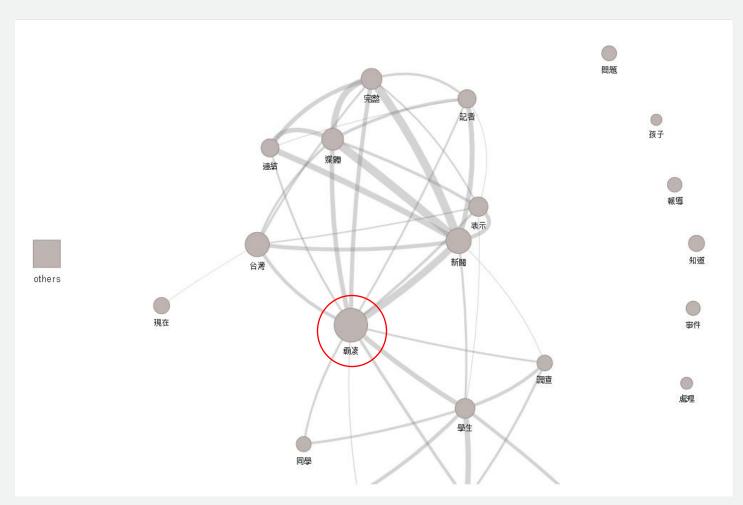


- 使用NTUSD字典
- 搭配使用否定詞與 加強詞

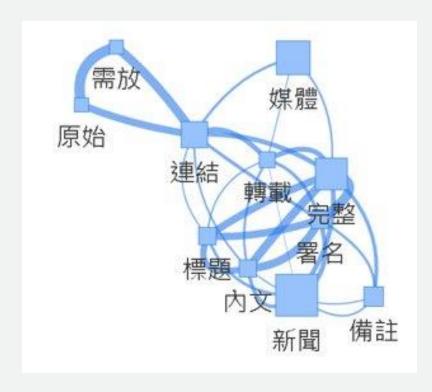
文辭及文件分析

- 運用 TF-IDF 來評估一個字詞對資料集(文件的 集合)或詞庫中一份文件的重要性
- 利用 DTM 獲得各文件裡面出現該字詞的頻率 (詞頻)

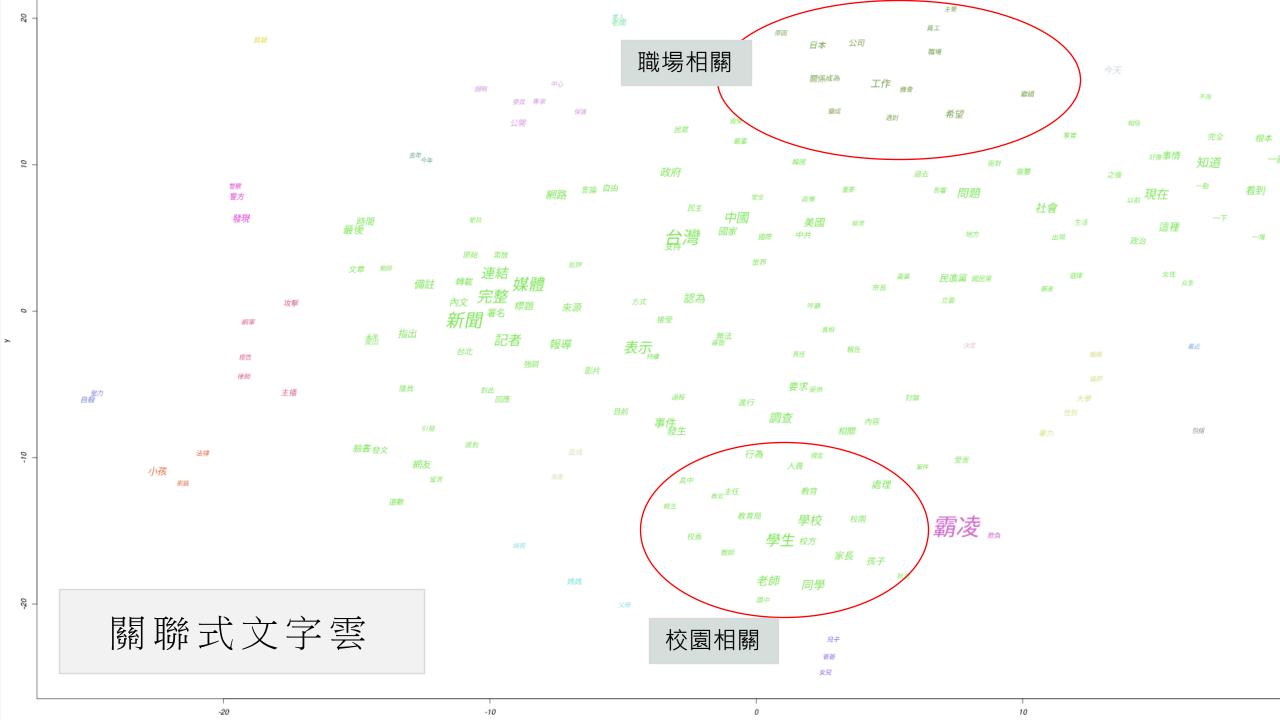




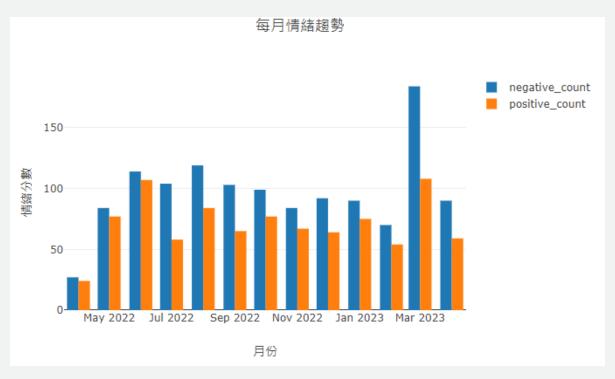
單中心網路圖

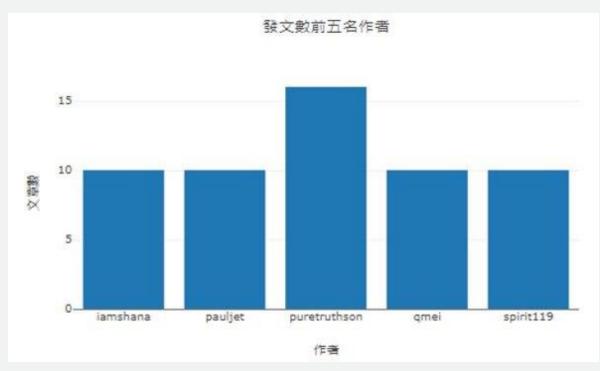


字詞網路圖



視覺化圖表分析

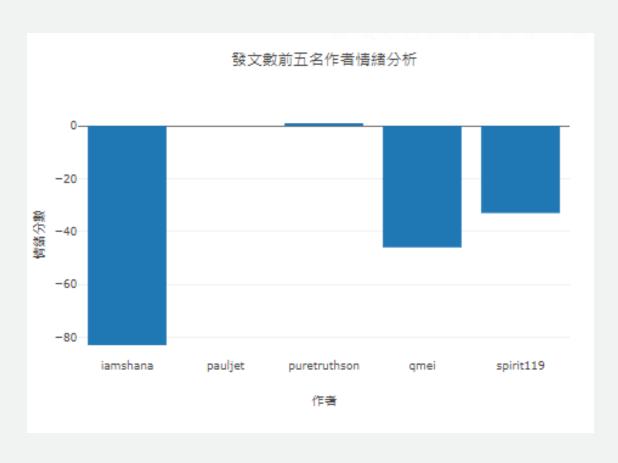




圖(一)每月情緒趨勢

圖(二)發文數前五名作者的文章情緒分數

視覺化圖表分析





圖(三)發文數前五名作者情緒分析

圖(四)每日情緒分數趨勢

視覺化圖表分析





圖(五)本議題文字雲

圖(六)詞性標記文字 雲

小結

- 1. 依Lexicon based情緒分析圖可觀察到,在正面的詞彙方面,「希望」出現的數量最高,其意義在於對於某種現況所期許的改變。而在負面字詞部分,「問題」這個字詞的貢獻度最高,霸凌本身就是以一種問題的形式來存在,當我們觀察到問題存在,並明白人們的關注及其意義的共識,這就會是需要得到解決的議題。
- 2. 從 Corpus based 情緒分析之結果來看,負面情緒的文章仍多於正面情緒的文章,由此可知在霸凌議題的討論上,負面情緒更容易被發文者表達出來,而且這些情緒更容易被鄉民關注及討論,因為這是一個嚴重的社會議題,通常使受害者痛苦及造成傷害,引起發文者憤怒、失望的負面情緒,進而激起許多人情感共鳴的反應。
- 3. 依LIWC情緒分析之結果來看,整體而言負向詞彙比例高於正向詞彙,推測其原因在於霸凌事件本身對社會對群體所帶來的大量負面效應,其中「問題」、「攻擊」、「欺負」等的負面詞彙出現較多,因霸凌本身即是一種社會問題,所伴隨著攻擊、欺負、傷害的行為也是由此問題而生,因此討論這些衍生行為亦會充斥於各個討論版上。

小結

4. 由 NTUSD 字典所獲得的分析結果,可以看出希「希望」一詞也在前三名字詞中出現,而「問題」一詞也依然佔據負面情緒詞彙中的榜首。

- 5. 使用進階字典情緒分析,並將選擇NTUSD字典使用,與前述單獨使用NTUSD分析有所不同的是,啟用否定詞與加強詞的解析,將「不」這樣的否定詞,與「非常」這樣的加強詞納入考量。
- 6. 經過 DTM 與關聯式文字雲的產出結果,可以看到工作、公司、職場這幾個關鍵字屬於同一群組,學生、校方、學校、老師、同學、家長則是較具集在一起,同屬一個討論範疇。

小結

7. 圖一、每月情緒趨勢:圖中呈現每月正面與負面情緒分數,每月的負面情緒分數皆高於正面情緒分數,且2023年3月差異最為顯著。

圖二、發文數前五名作者:提取文章總數最多前五名作者,了解對於霸凌議題關注度最高的作者 分別是哪幾位。

圖三、發文數前五名作者情緒分析:發文數最多前5名作者的情緒分析皆為負面分數,大致能夠推 論此議題的高度關注者情緒大多都是負面的。

圖四、每日情緒分數趨勢:將指定區間內每日關於霸凌的文章情緒分數進行趨勢分析,發現每日的情緒起伏跌宕,但偏向負面居多,可推知關於霸凌的議題的情緒分數大多是偏向負面的,這樣的結果也與預期相符:大多數發文對於霸凌本身有較多不滿的情緒用詞。

圖五、本議題文字雲:經過停用詞的清理後,執行詞頻計算並生成視覺化的文字雲,有助於更直觀 找出與關鍵字相關,出現數量較多的字詞。

圖六、詞性標記文字雲:可以讓我們直觀的看出本議題中,出現較多的是名詞以及動詞,結合圖五 觀察,可以再進一步去分析有些字詞,例如霸凌、攻擊等,在標記中是標為何種詞性,進而改善分 析結果。