Assessment 3: WebCrawler and NLP System

By: Jagraj Gill

Date Submitted: 28 Aug 22

**Overview**

The aim of this study is to deconstruct movie scripts that are associated with Oscar nominated and winning films via a subfield of sentiment analysis known as emotion detection (ED), whose primary goal is to extract the perceived emotions of textual data. The motivation behind this study is to better understand the commonalities and differences between the perceived emotions in these screenplays via statistical analysis and machine learning (ML), to determine what, if any correlations suggest as to what comprises an award-winning narrative. Current applications include hate speech detection, emotion retrieval from suicide notes, multimedia tagging, and analyzing e-commerce review sentiments (Yang et al., 2012; Wang et al., 2019; Allouch et al., 2018). Additionally, as noted *Figure 1* below, the ratio of text-based emotion detection is a limited study as compared to emotion detection in general, as seen in both IEEE Explore, and the Scopus sample databases (Acheampong et al., 2020).

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 1: The IEEE Explore Database is a digital repository that provides technical, journalistic, and conference-related materials related to computer science, electrical engineering, and electronics, containing over 5.7 million individual items (“Advancing Technology”, 2022). The Scopus Database is a provider of citation indices and abstracts pertaining to academic books, journals, and conference proceedings worldwide, containing over 39,000 titles (“Research Intelligence”, 2020). The increase in the frequency of research output in ED in multimodal forms such as speech, body language, and facial expressions as compared to text-based ED from 2010-2020 is evident in both academic databases.

Furthermore, the disparity in text-based ED is outlined mathematically as

 (1)

where *T* is the set of text to which the emotions are to be derived, *A* is the set of authors of *T*, *E* is the set of all emotions and *r* is a function that describes the emotions *E* of author *A* from text *T*. (Kao et al, 2009). The ED from text problem is derived from linking the input text to the actual projected emotion that leads the author to pursue specific artistic avenues in writing. Subsets of both *T* and *E* become non-trivial as they are either loosely defined, ambiguous, subject to language evolution, and finally, there are no set relations or classifications that encompass all human emotions.

When undertaking text-based ED, it is vital to define the model of emotion for use, as described in *Text-based emotion detection: Advances, challenges, and opportunities* by Anchempong et al. (2020). There are two primary models used in ED: discrete emotion, and dimensional emotion models. Discrete emotion models involve bucketing emotions into discrete categories, while dimensional emotion models assume that emotions are not independent, and are related to one another, hence should be mapped in a spacial format. There exist unidimensional and multidimensional models that are utilized to depict the degree of relation between emotions, and they typically reflect on two behavioral states, good, and bad. Both models are affected by relative degrees of their occurrences; however multidimensional models are more prominent and frequently used.

The two commonly used discrete emotion models include the Paul Ekman and Robert Plutchik models. The Paul Ekman model distinguishes emotions based off of six basic categories. This theory hypothesises that there exist six fundamental emotions that originate from separate neural networks in the brain, and result from how one perceives an event, hence emotions are manifested independently of one another. The six emotions are happiness, sadness, anger, disgust, surprise and fear. Furthermore, this theory states that more complex emotions may be derived from the synergy of the fundamental emotions.

The Robert Plutchik model suggests that there exist few primary emotions which take place as opposing pairs, and produce complex emotions by combination, as shown in *Figure 2* below. He adds two emotions to the Ekman model, acceptation, and anticipation. Plutchik postulates that for each emotion, there exists various degrees of intensities that occur because of how events are formulated by the experiencer.

Additionally, there exists the Orton, Clore, and Collins model, which opposed both Plutchik and Ekman’s basic emotion theory, proposing that emotions arise from the way they were perceived by the individual, and by the degree of their intensity. All three may be implemented in ED system design, but the OCC model allows for a broad spectrum of emotion representation, with 22 in total.

The primitive dimensional emotional model is presented by Russell as a two-dimensional representation known as the circumplex of affect, shown in *Figure 2* below. The model separates emotions in the arousal valence domains, with the arousal splitting the emotions by activations and deactivations, where the valence differentiates emotions by pleasantness and unpleasantness. In this model, emotions are not independent, but related. Discrete models have been accepted for their emotional classification capabilities because they are simple, but are not exhaustive of the full range of emotional classes available. They can map the intensities or degrees of occurrences as compared to dimensional models. Dimensional models are suggested for projects that involve comparing similarities in emotions.

Diagram

Description automatically generated Diagram

Description automatically generated

Figure 2: Plutchik’s wheel of emotions and Russell’s circumplex model of affects.

There are three primary techniques that are explored for text-based emotion detection, rule construction, machine learning, and a hybrid approach. The rules-based approach uses grammatical and logical rules to adhere to for ED. This method uses both keyword recognition and lexical affinity. Keyword recognition is reliant on emotion dictionaries or lexicons, some of them being Wordnet-Affect, EmoSenticNet, and the National Research Council of Canada lexicon. The emotion lexicons have emotion key words where the occurrences of them in text occurs at the sentence level. Once the keyword is determined, the sentence is labeled. A challenge that will be faced using this method will be the constraint imposed by the number of available emotions in the dictionary, navigation of word sense ambiguity, and processing of linguistic information when sentences are short or sparse of key words. To augment these limitations, the lexical affinity method may be explored, as random emotional words are assigned a probabilistic affinity on to of identifying keywords.

An example of challenges in the rule-based approach was set out in *Emotion and Sentiment Analysis from a Film Script: A Case Study,* by Yu et al. (2017), whose aim was to examine the emotions in the thriller movie script *A Hard Day*, as laid out by Plutchiks wheel of emotions. Manual sentiment (positive, negative, and neutral) and intensity tagging were conducted by undergraduate students, and was compared with the results of the python-based NLTK (Natural Language ToolKit) package VADER (Valence Aware Dictionary and sEntiment Reasoner) 2.5. The result was a 42.6% matching rate when comparing both ED and intensity measures after averaging the negative and positive emotions. The barrier faced in this study was analogous to finding a universal relation function *r* for subsets of text and authors as outlined above in Kao et al. (2009). Future aspirations of the researchers included adding more human taggers and adopting ML algorithms.

ML-based models have been widely implemented for text-based classifications, where supervised machine learning models have performed better in emotion detection rates as compared to traditional unsupervised machine learning (TUML).

The more widely explored unsupervised methods have been traditional in nature (i.e., Support Vector Machines, Naïve Bayes), which have proven to be less robust, and do not explicitly extract semantic information from texts which is needed for effective emotion detection. One avenue of approach that may serve as an intermediary step This method has not been explicitly discussed in the current literature. The current science supports deep learning (DL) approaches as layers can extract, or learn, the intrinsic properties that text carries, hence some models suggested by the initial paper outline DL outperforming TUML.

The final avenue that is typically considered is a hybrid approach, where rule construction and ML will deliver a unified model which takes in the strengths of each and mitigates their limitations. Acheampong et al. (2020) also indicates in its survey that this results in “satisfactory results” referring to TUML and rule construction methods, and points to a void in exploring effective approaches with deep learning models, as this is the latest advancement and optimal modalities have not been established.

To expand research in the aforementioned domains and their

**Domains**

Four websites were considered candidates for web scraping: Simply Scripts (<https://www.simplyscripts.com>), Script-O-Rama (<http://www.script-o-rama.com>), The Internet Movie Script DataBase (<https://imsdb.com>), and Scripts Plug (<https://www.scriptslug.com>). One reason for utilizing these databases was significant script volume pertaining to Oscar-winning screenplays and similarly for unremarkable titles. All websites were HTML (HyperText Markup Language) and XML (Extensible Markup Language)-based hence there were multiple options available for parsing their data: Python’s *request* module, BeautifulSoup, Scrapy, and Selenium. Determining the avenue that will be more efficient and beneficial to use will come down to the structure of the website of interest (i.e., Selenium handles core JAVA applications well), whereas in terms of speed and efficiency, Scrapy would be more beneficial as it supports multithreading and pre-build requests and parsing modules. BeautifulSoup is regarded for its comprehensive documentation and ease-of-interpretation, while the *requests* module is known for performing common HTTP requests (i.e., GET, POST, PUT, PATCH, and DELETE) in a low-code manner.

Simply Scripts

There were no limitations faced

Coverage,limitations of wc, data on the site, and any copyright/legal concerns

**WebCrawler Workflow**

**Data Wrangling**

**Data Summarization**

**Machine Learning Structure**

**Evaluation of Model**

**Discussion**

This study will consider the influence and popularity terms to be synonymous to one another. It is important to create this ambiguity as influence could be subjective to the audience, and popularity may be determined by uncontrolled factors such as availability, outreach, population size, and interpretability barriers. Although some modeling techniques such as a Bayesian fixed effects model may be able to find a signal for such attributes, this is beyond the scope of complexity that is targeted with this investigation, and does not directly target the aim and motivation.

**Conclusion**

**References**

Acheampong, F. A., Nunoo-Mensah, H., & Wenyu, C. (2020, May 28). *Text-based emotion detection: Advances, challenges, and opportunities*. Wiley Online Library. Retrieved August 21, 2022, from <https://onlinelibrary.wiley.com/doi/full/10.1002/eng2.12189>

Yu, H.-Y., Kim, M.-H., & Bae, B.-C. (2017, December 25). *Emotion and sentiment analysis from a film script: A case study*. Journal of Digital Contents Society. Retrieved August 20, 2022, from <http://koreascience.or.kr/article/JAKO201707356128225.page>

*Research intelligence - Scopus Content Coverage Guide*. Elsevier. (2020). Retrieved August 21, 2022, from <https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf>

IEEE. (2022, August 21). *Advancing Technology for Humanity*. IEEE Xplore. Retrieved August 21, 2022, from <https://ieeexplore.ieee.org/Xplore/home.jsp>

E. C. -C. Kao, C. -C. Liu, T. -H. Yang, C. -T. Hsieh and V. -W. Soo, "Towards Text-based Emotion Detection A Survey and Possible Improvements," 2009 International Conference on Information Management and Engineering, 2009, pp. 70-74, doi: 10.1109/ICIME.2009.113.

Yang Hui, Willis Alistair, De Roeck Anne, Nuseibeh Bashar. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes. Biomedical Informatics Insights. 2012;5(1):BII.S8948. <http://dx.doi.org/10.4137/bii.s8948>.

Wang Shangfei, Peng Guozhu, Zheng Zhuangqiang, Xu Zhiwei. Capturing Emotion Distribution for Multimedia Emotion Tagging. IEEE Transactions on Affective Computing. 2019;1–11. <http://dx.doi.org/10.1109/taffc.2019.2900240>

Allouch M, Azaria A, Azoulay R, Ben-Izchak E, Zwilling M, Zachor DA. Automatic detection of insulting sentences in conversation. Paper presented at: Proceedings of the 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE); 2018:1-4; IEEE.