Strengths and Limitations of Artificial Intelligence

Machine Learning and Natural Language Processing of Social Media

Student Name (Student ID)

Faculty of Science, Engineering, and Technology,
Swinburne University

Hawthorn, Australia

email@student.swin.edu.au

I. Introduction

A panel of ACM Turing award recipients indicated that there have been major recent break throughs in the field of Artificial Intelligence (AI) regarding machine learning, as well as speech and visual object recognition [1]. The panel further suggested that promising areas of AI application were in health care, self-driving cars, robotics, and natural language processing (NLP). The panel conceded however that it was likely to be a long time before AI could be considered as truly 'intelligent'.

The comments of the ACM panel suggest that AI has both strengths and limitations, however without knowing more about AI, it is difficult to appreciate what these may be. The purpose of the current review is to provide some insight into the strengths and limitations of AI by reviewing academic literature from the past few years on machine learning (ML) and NLP. It achieves this by focusing on the topic of 'sentiment analysis' – a form of NLP that may involve ML.

The review begins by presenting a general overview of sentiment analysis and its application to social media. This overview highlights the importance of specialised dictionaries for NLP, training sets for ML, and noise issues related to social media data. A review of recent sentiment analysis articles is then conducted highlighting a variety of issues that appear to impact on analysis accuracy. The review then discusses figurative language detection as an example of the kinds of challenges and complexities that may be experienced in the field of AI.

II. Review Method

To get an overview of the domain, a survey of articles was conducted on the database Scopus (c.f. [2]). The search terms were "natural language" and "machine learning" and the results were limited to 'computer science', 'articles' and 'reviews' only, and to publications from 2017 onwards. The abstracts of the resulting 74 matches were read and categorised. Seven articles were omitted for being off-topic. Figure 1 displays the categories found.

Representative subcategories were selected for review. 'Sentiment analysis' was selected from within the NLP category and 'social media' as the domain application. These

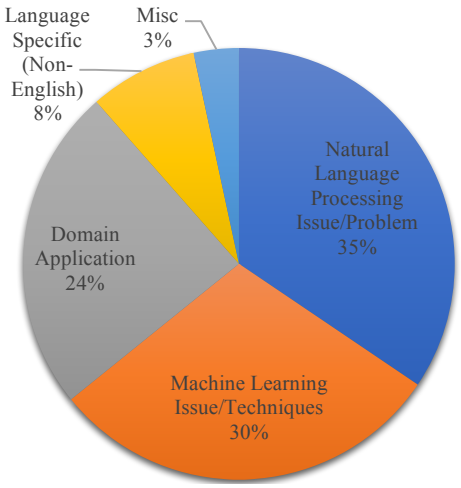

Fig. 1. Survey of Natural Language and Machine Learning Articles by Category on Scopus: Jan-Aug 2017.

were chosen as they represented substantial and coherent research topics within the domain, whereas many others were more diverse and didn't attract multiple researchers. Using these categories, a new search was conducted on Scopus, and all articles fitting the categories from 2015 to present (August 2017) were collected.

The 17 articles collected were categorised again, and 6 articles were selected for comprehensive review. The topics selected were; 'sentiment analysis technique comparisons' (3 articles), and 'figurative language detection' (3 articles). These were chosen as they seemed representative of the strengths and limitations of AI.

III. Review of Relevant Works

A. Domain Overview

Sentiment analysis is the categorisation of documents into positive and negative opinions [2, 3]. This is generally achieved via the use of special dictionaries called lexicons that categorise words into positive, negative, or neutral sentiments [2, 3]. Documents are parsed through lexicons and the resulting polarity data is then used as the input for ML algorithms [2, 3]. The ML algorithms are then trained to find relationships between the polarity data and the desired

outcome categories using a corpus of documents that have already been classified (i.e., annotated) [3]. The ML classifiers are then used to categorise further documents into the learned categories [2, 3].

Lexicon parsing and ML techniques may also be used independently [2]. For example, a lexicon may be used to determine the polarity of words in a document, and the result of the sum used for classification [4]. It is also possible to train ML algorithms to find interrelationships between words and categories using an annotated training set only, without the use of a lexicon [4, 5].

Social media data introduces unique challenges for sentiment analysis. A recent review of the analysis of micro-blogs (i.e., Twitter) showed that although social media presents a rich source of data, the use of informal and platform dependent language creates additional sources of noise [6]. For example, social media often involves heavy use of abbreviations, slang, poor grammar and spelling, and user generated annotations in the form of emoticons and hashtags [6]. The presence of such features requires analysts to determine normalisation strategies for the data and make decisions about what to include [6].

B. Sentiment Analysis

The first article reviewed presented a case study for the use of social media sentiment analysis by governments [7]. The authors compared a lexicon based approach, a ML approach, and an established lexicon-ML approach called SentiStrength. The authors asserted that there was consistency between approaches in the results, thus implying the overall validity of sentiment analysis for the domain. The reported results however, indicated a significant statistical difference between the approaches with a small-medium effect size ($\eta^2 = 0.21$, $p = 0.001$). Furthermore, no measures of accuracy for the approaches was presented, meaning that their validity was assumed by the (statistically failed) cross-validation. If it is assumed that the sentiment analysis was accurate, the article showed how it may be used to track public opinion over time. However, as subsequently reviewed articles show, accuracy can be quite low in some situations.

The next article compared sentiment analysis for whole documents and for specific topics [4]. Whole documents ignore what topic(s) may be present within it, and specific topics are extracted from within whole documents. The authors used a combined lexicon-ML approach for classification with many additional lexical analyses that went beyond simple polarity classification. They compared this model to a simple lexicon model and a ML model on various established document corpuses, mainly from product review sites but also from Twitter. The corpuses used ranged from small (5,000 documents) to very large (50,000 documents). The results indicated that whole document analysis performed best with 80-90% accuracy on large, focused datasets. Topic analysis was less accurate at 70-80% on the same data. On data sources that involved more diversity however, accuracy was much lower; e.g., one Twitter data source was 60-65% accurate for whole document analysis. Whilst the authors argued for the superiority of their combined lexicon-ML model, the results indicated that it was only about 1% more accurate than simple ML. The lexicon only approach however, performed about 15% worse than the other models.

The final article on sentiment analysis involved the compilation of a corpus of tweets, as well as a comparison of three ML algorithms [5]. This work did not involve a lexicon. The corpus consisted of 34,634 tweets that were annotated by emoticons. The ML algorithms that were used were Support Vector Machine (SVM), Naïve Bayes (NB), and logistic regression (LR). The results showed that SVM had superior accuracy at 73-74%, which was about 8-12% better than the other algorithms. Whilst the authors also tested a variety of input weighting models for the data, these had a negligible effect. Whilst the results were reasonably robust in comparison to the previously reviewed article [4], the use of emoticons as annotations may well have introduced a source of noise to the data, as there is unlikely to be any consistency between people's reasons for using them.

Overall, the review of sentiment analysis highlighted some principal issues. Notably, the accuracy of classifiers can range widely, and cannot be automatically assumed to be a valid. Next, whilst lexicons appear to improve basic ML techniques, the effect of them is largely trivial. More importantly, the quality and size of the dataset used to train and run ML classifiers on appears to have a profound impact on their efficacy.

C. Figurative Language Detection

Figurative language detection is a challenge in sentiment analysis, as its presence in the form of satire [8], irony, and sarcasm [9, 10] can reverse the polarity of word meanings. Naturally this is considered as a potential source of noise that can reduce the accuracy of classifiers.

The first article reviewed on figurative language focused on satire [8]. This research involved creating a corpus of 10,000 tweets from Twitter collected from legitimate and satirical news organisations. A combined lexicon-ML approach was used, and the lexicon involved many categories beyond simple polarity. Three ML algorithms were trained and compared on their ability to detect satire, these were; SVM, Bayes-Net, and J48 (a decision tree model). The results showed that SVM performed best with accuracies between 84-86%, which was 5-10% more accurate than the others. Compared with the other studies reviewed, this was a robust result. Following the idea that dataset quality has an impact on performance; the fact that the dataset was annotated using clearly-defined sources (real vs satirical news organisations) goes further to reinforce this notion.

The second article reviewed investigated the categorisation of sarcasm, irony, and metaphor [10]. The authors of this study used two different lexical categorisation techniques (polarity and cluster analysis) with a hybrid SVM-decision tree ML classifier. The ML was trained on a corpus 8,000 tweets that were annotated via a crowd sourcing service. It was then tested on a small corpus of 927 tweets annotated as ironic, sarcastic, or metaphorical, and a larger corpus of 4,000 tweets which also contained non-figurative examples. The cluster analysis technique produced the most

accurate performance on the smaller, focused dataset with an accuracy of 78%. On the larger dataset, performance was poor with only 55% accuracy. These results further reinforce the notion of the quality of the dataset to ML performance. They go further however to indicate that the training set and test set need to be of a similar composition. The training set in this study only involved irony, sarcasm, and metaphorical tweets and not neutral ones as it was later tested on.

The last article focused on sarcasm in conjunction with regular sentiment analysis [9]. This article created a small corpus of 2,700 tweets and used hashtags as annotations (e.g., #sarcasm, #happy, #sad). Three ML algorithms that used a sentiment lexicon were compared. The ML algorithms were; SVM, NB, and LR. The SVM model outperformed the other algorithms by between 2-4% on accuracy, however only achieved 60% accuracy overall. The authors then increased the size of their corpus up to a total of 100,000 tweets, which resulted in improved accuracy of 67-73%. The authors also compared these performances to those of humans, and found that they were on par. However, the human annotators had low agreement on classifications (50%), but on those that they did agree on they outperformed the classifiers substantially (87% compared to 60%). Whilst the authors argued that sarcasm detection was difficult for humans and ML alike, the use of hashtags to annotate the corpus (and to derive accuracy ratings) is questionable as highlighted by the lack of agreement between the human judges.

Overall the review of figurative language detection further reinforced the notion that clean, clearly defined, and representative training datasets for ML are important for performance. It also highlighted the superiority of the SVM classifier in the social media, NLP domain.

IV. Conclusions

This review has highlighted some issues which may potentially extrapolate to other AI research and application domains. Most notably, it is clear from the literature reviewed the ML performs best in coherent and constrained conditions. Training data needs to be rigorously formed with clear and consistent classification definitions used for the annotation. Furthermore, such training material needs to be a good sample representation of the actual domain of subsequent application. Taken together these ideas suggest that AI is best suited to coarse grained tasks within well-defined problem spaces. Since size matters for training set data, finding methods to create suitable training materials is an important challenge to resolve for those working in the field.

More specifically to the field of NLP and sentiment analysis, it is also clear that SVM represents the current state-of-the-art for ML classification algorithms. The present review suggests that there is only a negligible advantage to utilising lexicons and other lexical analysis techniques, contrary to their apparent popularity. Instead rigorous and consistently annotated training sets for ML have been shown to be more important.

Word Count: 2,007

References

[1] CACM Staff, "Artificial intelligence," *Commun. ACM,* vol. 60, no. 2, pp. 10-11, 2017.

[2] M. Injadat, F. Salo, and A. B. Nassif, "Data mining techniques in social media: A survey," *Neurocomputing,* vol. 214, pp. 654-670, 2016/11/19/ 2016.

[3] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion,* vol. 36, pp. 10-25, 2017/07/01/ 2017.

[4] V. Hangya and R. Farkas, "A comparative empirical study on social media sentiment analysis over various genres and languages," *Artificial Intelligence Review,* Article vol. 47, no. 4, pp. 485-505, 2017.

[5] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and R. Mitkov, "Polarity classification for Spanish tweets using the COST corpus," *Journal of Information Science,* Article vol. 41, no. 3, pp. 263-272, 2015.

[6] R. Srivastava and M. P. S. Bhatia, "Challenges with sentiment analysis of on-line micro-texts," *International Journal of Intelligent Systems and Applications,* Article vol. 9, no. 7, pp. 31-40, 2017.

[7] H. M. Chen, P. C. Franks, and L. Evans, "Exploring government uses of social media through Twitter sentiment analysis," *Journal of Digital Information Management,* Article vol. 14, no. 5, pp. 290-301, 2016.

[8] M. D. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodriguez-García, R. Valencia-García, and G. Alor-Hernández, "Automatic detection of satire in Twitter: A psycholinguistic-based approach," *Knowledge-Based Systems,* Article vol. 128, pp. 20-33, 2017.

[9] S. Muresan, R. Gonzalez-Ibanez, D. Ghosh, and N. Wacholder, "Identification of nonliteral language in social media: A case study on sarcasm," *Journal of the Association for Information Science and Technology,* Article vol. 67, no. 11, pp. 2725-2737, 2016.

[10] H. L. Nguyen and J. E. Jung, "Statistical approach for figurative sentiment analysis on Social Networking Services: a case study on Twitter," *Multimedia Tools and Applications,* Article vol. 76, no. 6, pp. 8901-8914, 2017.