Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), pages 38–42
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Does History Matter?
Using Narrative Context to Predict the Trajectory of Sentence Sentiment

Liam Watson, Anna Jurek-Loughrey, Barry Devereux, Brian Murphy
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Northern Ireland
{lwatson11, a.jurek, b.devereux, brian.murphy}@qub.ac.uk

Abstract

While there is a rich literature on the tracking of sentiment and emotion in texts, modelling the emotional trajectory of longer narratives, such as literary texts, poses new challenges. Previous work in the area of sentiment analysis has focused on using information from within a sentence to predict a valence value for that sentence. We propose to explore the influence of previous sentences on the sentiment of a given sentence. In particular, we investigate whether information present in a history of previous sentences can be used to predict a valence value for the following sentence. We explored both linear and non-linear models applied with a range of different feature combinations. We also looked at different context history sizes to determine what range of previous sentence context was the most informative for our models. We establish a linear relationship between sentence context history and the valence value of the current sentence and demonstrate that sentences in closer proximity to the target sentence are more informative. We show that the inclusion of semantic word embeddings further enriches our model predictions.

Keywords: emotion, sentiment, valence, narrative fiction, word embeddings

1. Introduction

The experience of emotion plays a major role in the way people understand and engage with stories. In works of literary fiction, it is the affective trajectory of the story (the emotional journey that the reader is taken on) that propels the plot forward. People read stories because they are emotionally invested in the fates of the characters. In Natural Language Processing (NLP), there is a rich literature on using lexical, semantic and structural information to infer an emotional tag or value for sentences and short passages (Pang et al., 2008; Cambria, 2016; Mohammad, 2016; Liu, 2010). However, modelling the emotional trajectory of narratives poses new challenges – a model must be able to account for both the long distance effects of previous discourse on the reader, and the contextually subtle ways in which the high-level information conveyed by a text can influence the reader's emotional state.

The field of sentiment analysis (i.e. the task of "automatically determining valence, emotions, and other affectual states from text" (Mohammad, 2016)) has begun to answer the question of how we can evaluate the emotional content of text, particularly with regard to commercial domains and social media. For example, work on sentiment analysis has focused on product or movie reviews (Mohammad, 2016; Liu, 2010; Socher et al., 2013; Tai et al., 2015) or on the analysis of twitter feeds (Liu, 2010; Zimbra et al., 2018). Recent work using deep learning, and in particular recurrent neural networks (RNN) such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), and Transformer networks (Vaswani et al., 2017) has facilitated a significant increase in the performance of sentiment classification of texts and, given the ability of such networks to represent information over long sequences (Socher et al., 2013; Tai et al., 2015; Jiang et al., 2019), they show particular promise for modelling high-level properties of natural discourse, such as literary texts. Most of the work on sentiment analysis makes use of large, readily available corpora of labelled data, which contain short samples of text (e.g. tweets or movie reviews) and associated explicit rating values (e.g. 5-star rating systems for movie and product reviews, or emoticons or hashtags used to summarise or emphasise the emotional content of a tweet (Liu, 2010; Mohammad, 2016; Socher et al., 2013; Tai et al., 2015). However, no large dataset of literary text annotated for emotional content exists, and so in this study we start by developing a method which can learn to predict the emotional content at a particular point in a story given the preceding context and existing word-level resources (such as hand-tailored sentiment dictionaries, and corpus-derived word-embeddings). In particular, in order to determine how the sentiment of the text changes over time we must evaluate the sentiment of each new sentence as it arises within the context of the text that has come before. Our approach conceives the problem of modelling the emotional trajectory of narrative as consisting of two distinct questions:

1. Can the sentiment of a given sentence be determined by a previous history of sentences?
2. How much history should be included to be optimally informative?

We focus on modelling emotional valence at the sentence level. Explicitly, we model the valence of any given sentence in a sequence of sentences making up a narrative using the preceding context. We explore various sizes of sentence history context window and the effects of incorporating semantic information through the inclusion of pretrained word embeddings of various dimensions.

To our knowledge, very little previous work has directly examined the influence of sentence history on the current sentence's valence as we do in this paper. Jockers (2015) takes a simple sum of word valences as representative of sentence valence and then employs a number of different smoothing functions to allow for the effects of history. Whissell (2010) takes a mean of all word valence values as representative of the valence value for different chunks of text (e.g. sentence, paragraph, and chapter-level chunks). In this work, we choose sentence-level sentiment as the best basic unit of measurement for emotional content. We model sentence-level valence using a lexicon of sentiment (Whissell, 2010), where the sentence-level valence is estimated as the mean of the sentence's word valences as found in the lexicon. While we are aware that a sentence valence rating based on a mean of the constituent word ratings taken from a lexicon is not state-of-the-art in sentiment analysis, the approach is validated by work in psychology (Whissell, 2010; Whissell, 2003; Bestgen, 1994) and offers a computationally inexpensive way to begin this exploratory work, in the absence of large labelled datasets.

2. Related work

Most work in the field of sentiment analysis has focused on product reviews, tweets, and emails, and has been focused on determining opinions towards certain targets (e.g. the new iPhone, or President Obama) (Mohammad, 2016; Liu, 2010; Mohammad et al., 2013). Liu (2010) surveys the field of sentiment analysis with a focus on opinion mining — determining users opinions about goods or services by analyzing reviews. Mohammad et al. (2013) trained two SVM classifiers for two different sentiment tasks; the first of these was a message level sentiment prediction task and the second a term-level task. They achieved state-of-the-art performance on both tasks using two lexicons generated from tweets (the first using tweets with sentiment hashtags to generate the lexicon, the second using tweets with emoticons). The use of such lexicons of affect, where each entry is annotated with a valence value, is commonplace in sentiment analysis. As well being automatically generated, as in the tweet lexicons (Mohammad, 2016), lexicons may also be created by human annotation (usually gathered using online tools such as Mechanical Turk).

There are several prominent sentiment lexicons that differ in their contents and methods of compilation. The NRC Emotion Lexicon, known as Emolex (Mohammad and Turney, 2010), is a list of 14,182 English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The terms in EmoLex are carefully chosen to include some of the most frequent English nouns, verbs, adjectives, and adverbs. The Opinion Lexicon (Liu et al., 2005) consists of a list of 6800 positive and negative sentiment words. This lexicon only consists of words believed to be associated with either polarity and does not contain any neutral words. AFINN (Nielsen, 2011) is a list of English words rated for valence on a scale of -5 (negative) to +5(positive). The words were manually labeled by Finn Årup Nielsen (the author) in 2009-2011. There are two versions of this lexicon — AFINN-96 (1468 unique words and phrases) and AFINN-111 (the newest version with 2477 words and phrases). There are also lexicons available from studies on emotion in psychology, most notably the Revised Dictionary of Affect in Language (DAL) (Whissell, 2010). Whissell's DAL consists of 8742 English words which have been rated for their activation, evaluation and imagery. Each of these dimensions was rated along a three point scale: (1) Unpleasant, (2) In between, (3) Pleasant; (1) Passive, (2) In between, (3) Active; (1) Hard to imaging, (2) In between, (3) Easy to imagine. It was comprised of frequently occurring words in a number of sources including an established corpus of 1,000,000 words (Francis and Kucera, 1979), samples of writing generated by adolescents, and juvenile literature. When tested against a corpus of 350,000 English words gathered from many different sources, the DAL demonstrated a matching rate of 90%, suggesting that we can expect 9 out of every 10 words in any given English language text to have rating data in DAL (Whissell, 2009).

There is some work to demonstrate that there is a correlation between these lexical affective word ratings and subjective passage ratings (Bestgen, 1994; Whissell, 2003; Hsu et al., 2015). However, these studies have relied on carefully chosen text inputs and have avoided complicating issues such as negation and irony, etc., which are commonplace in natural discourse.

While there have been a few studies into emotion in literary texts (Bestgen, 1994; Mohammad, 2012; Whissell, 2003; Hsu et al., 2015), these have largely focused on detecting discrete emotions (love, anger, fear etc.) and centred almost exclusively on classifying texts (or sections of text) into these discrete groups. Mohammad (2012) compared the polarity and emotional word density (defined as the number of emotion words per X-words) of novels and fairy tales in English. Using the NRC Emotion lexicon, Mohammad and Turney (2010) labelled words in novels and fairy tales with polarity and discreet emotions such as joy, sadness, and so on. They then used an emotion analyser tool to make certain inferences from the data; for example, counting the instances of words related to particular emotions, and comparing the emotional distributions of different words across different genres. However, this work focused on discreet emotions (joy, anger, etc.) using associated emotion words, which can enlighten us in terms of literary criticism or text classification, summarization, etc., but which are not sufficient to help us to effectively model the emotion of a text in a way comparable to how a person experiences it over time as a story unfolds, or how it is constructed in the brain. Reagan et al. (2016) investigated the emotional arcs of narrative fiction using a sliding window of sentences.

What all of the aforementioned approaches have in common is that they consider the task of investigating valence and emotion in literature as a classification problem. The goal is to assign a given text or segment of text with a valence label which can then be used to derive some insight into the author's opinion regarding some product or issue, or to bring some quantitative insight to bear on studies in literary criticism. In this study, in contrast, we aim to model the changing experience of emotion during the course of reading a text. For this reason we frame the problem as a regression task, where we aim to predict a real number (measuring the degree of positive or negative emotion) for each sentence in the sequence of sentences making up the narrative.

3. Methodology

We aim to predict the valence of each sentence using information extracted from the history preceding that sentence. For this purpose, we train machine learning models that assign an emotion value to each sentence given information available in the preceding context. There are three key challenges that need to be addressed. First, identifying the features of the preceding context that are relevant to this sentence-by-sentence valence assignment task. Second, identifying what size of context history is most informative. And third, determining the type of machine learning model which performs best in predicting these sentence valences. As a first step, we investigate the degree to which the relationship between current sentence valence and sentence context history information can be modelled using linear methods. We apply two models to this task — linear regression and a linear support vector regressor. In the second part of the study, we investigate whether the application of non-linear methods to the same feature sets can better model the relationship between the sentence context history and the current sentence valence. We implement these non-linear models using a random forest regressor.

To train these models we explore a number of different feature combinations, to determine which kinds of information are most important for predicting sentence-level valence. We explore the scope of context relevant to inferring sentence valence, investigating different sizes of sentence context history and a variety of feature sets of different dimensionalities. This first stage of our study therefore focuses on the exploration of eighteen different feature sets combined in the following ways: (1) a history of sentence valence scores only (over a number of history window sizes, spanning 10, 50 and 100 sentences), and (2) a history of sentence valence combined with semantic information (i.e. pre-trained semantic word embeddings in the form of 50, 100, 200 and 300 dimension GloVe word embeddings (Pennington et al., 2014), and 300 dimension FastText word embeddings (trained on subword information) (Bojanowski et al., 2017) again over the same number of context history window sizes (10, 50 and 100 sentences). The 18 different feature set combinations investigated correspond to the rows of the results table below (Table 1).

4. Data and Resources

4.1. Text Used

Project Gutenberg (https://www.gutenberg.org/) provides access to thousands of public domain books (copyright expired) in plain-text format. We selected a corpus of 100 books (643,352 sentences) in total. We split these, by book, into 72 training texts (476,891 sentences, 74% of our corpus) and 28 test texts (166461 sentences, 26% of our corpus). The texts were split in this way to preserve the natural boundaries between books. These books were chosen as they represent pieces of literary fiction for children which would be well in common narrative techniques such as the use of irony, metaphors and imagery, and creative language. These are important features of literary language which can prove challenging for sentiment analysis systems based on a simple literal interpretation of sentences.

4.2. Lexicons and lexical embeddings

In training our models, we used information about emotional content derived from Whissell (1989)'s Dictionary of Affect in Language (the Revised DAL) (Whissell, 2010), discussed in Section 2. We generated sentence-by-sentence valence ratings for our target texts using the Whissell lexicon. The valence for each sentence is estimated by averaging over the valence values for the constituent words in the sentence. We then took these sentence-level valence ratings as the target values we hoped to predict.

5. Results

We explored three different machine learning models: Linear Regression, Linear Support Vector Regression and Random Forest Regression. The results from these models (R^2 values for predictions on the test set) are displayed in Table 1 below. We also present two figures which each illustrate different patterns observable from the data. Figure 1 illustrates the difference in performance of each of the machine learning models tested, across each of the different context windows. Figure 2 shows the difference in performance on each feature set across all of the models tested.

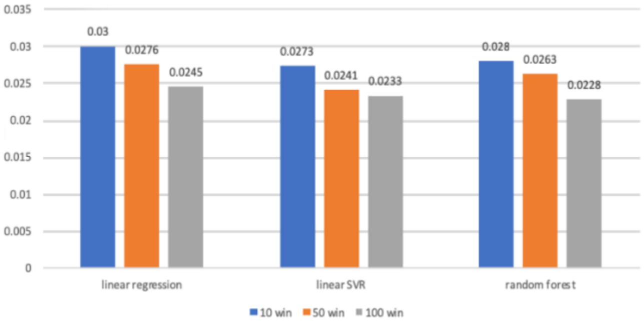

Figure 1: Performance (R^2 values on the test set) of all machine learning models across all context sizes, averaged over all feature sets.

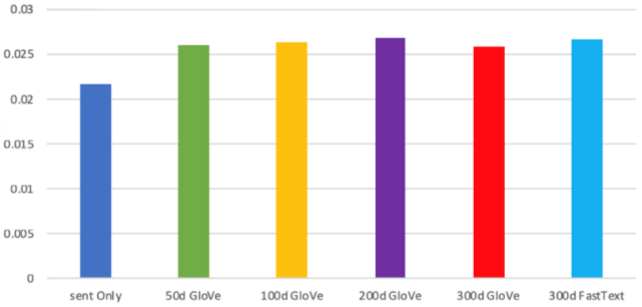

Figure 2: Contribution of each different feature combination to model performance; averaged over all model sets.

Feature Set	Context	Linear Regression	Linear SVR	Random Forest
Sentence Only	10	0.0210	0.0210	0.0236
	50	0.0215	0.0215	0.0226
	100	0.0215	0.0214	0.0213
50d GloVe	10	**0.0309**	**0.0312**	0.0295
	50	0.0304	0.0250	0.0285
	100	0.0280	0.0244	0.0261
100d GloVe	10	0.0309	0.0203	0.0294
	50	0.0302	0.0238	0.0267
	100	0.0274	0.0231	0.0256
200d GloVe	10	0.0302	0.0308	0.0291
	50	0.0288	0.0251	0.0267
	100	0.0259	0.0239	0.0214
300d GloVe	10	0.0288	0.0294	0.0283
	50	0.0273	0.0242	0.0261
	100	0.0235	0.0231	0.0211
300d FastText	10	0.0299	**0.0312**	**0.0310**
	50	0.0273	0.0249	0.0271
	100	0.0241	0.0237	0.0214

Table 1: Performance (R^2 values for predictions on the test set) of all machine learning models across all context sizes.

6. Discussion

Our study has focused on two central questions – firstly, to establish whether linear or non-linear methods are best suited to modelling this type of relationship and, secondly, to determine what kind of features extracted from the historical content are the most effective in training the machine learning models. This second question of finding an optimal feature set can be sub-divided into two smaller problems: (a) assessing whether the inclusion of semantic information in the form of pre-trained word-embeddings adds more relevant information to the model training, and (b) determining if there is an optimal size of sentence history context that should be included to generate the best predictions for each model.

From the results presented in Table 1, we can see that there is a small linear relationship between sentence valence history and the valence of the current sentence. This relationship is statistically significant at $p = 0.0001$. While these results clearly show that we have captured a real linear effect between valence history and current sentence valence, the magnitude of explained variance is small. The application of non-linear methods does not improve performance. However, we can discern an important pattern in these results regarding the influence of sentence history context on our model predictions. We can see from Table 1 that across all models and feature sets, the best results are generated using a sentence history context of 10 sentences, which confirms our intuition that sentences closer to the sentence being predicted should bear more on its valence value than sentences further back in the history. This information is summarised in Figure 1 where we have taken an average across all feature sets for each model to illustrate this trend.

Figure 2 depicts a summarisation of the relative contribution of each of the feature sets averaged across all of the models implemented and all of the context history sizes employed. We can see from this illustration that while all of the feature sets ultimately result in models which exhibit similar performance, in general, the inclusion of the semantic word embeddings does add slightly to the predictive power of the models.

7. Conclusions and Future Work

In this paper we proposed to investigate whether information present in a history of previous sentences can be used to predict a valence value for the following sentence in context. We explored both linear and non-linear methods and a range of different feature combinations. We also looked at different context history sizes to determine what range of previous sentences was most informative for our models. In conclusion, we have established a linear relationship between sentence context history and the valence value of the current sentence. We have demonstrated that the sentences in closer proximity to the target sentence are more informative. We have also shown that the inclusion of semantic word embeddings does seem to enrich our model predictions. We have therefore established a firm base for further explorations of valence in literature which should be characterised by further investigations of potentially optimally informative feature sets and the application of models capable of better capturing the complex, non-linearities inherent in literary text, such as LSTM artificial neural networks.

8. Bibliographical References

Bestgen, Y. (1994). Can emotional valence in stories be determined from words? *Cognition & Emotion*, 8(1):21–36.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Francis, W. N. and Kucera, H. (1979). *Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, USA.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, C.-T., Jacobs, A. M., Citron, F. M., and Conrad, M. (2015). The emotion potential of words and passages in reading harry potter–an fmri study. *Brain and language*, 142:96–114.

Jiang, M., Wu, J., Shi, X., and Zhang, M. (2019). Transformer based memory network for sentiment analysis of web comments. *IEEE Access*, 7:179942–179953.

Jockers, M. L. M. (2015). Revealing sentiment and plot arcs with the syuzhet package. (blog), february 2, 2015.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA. ACM.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June. Association for Computational Linguistics.

Mohammad, S., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Nielsen, F. A. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, et al., editors, *MSM*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98. CEUR-WS.org.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Whissell, C. M. (1989). The dictionary of affect in language. In *The measurement of emotions*, pages 113–131. Elsevier.

Whissell, C. (2003). Readers' opinions of romantic poetry are consistent with emotional measures based on the dictionary of affect in language. *Perceptual and motor skills*, 96(3):990–992.

Whissell, C. (2010). Whissell's dictionary of affect in language: Technical manual and user's guide. *Laurentian University*.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. (2018). The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29.

9. Language Resource References

Whissell, C. (2010). Whissell's dictionary of affect in language: Technical manual and user's guide. *Laurentian University*.