Long distance history influences sentence valence in narrative texts

Liam Watson*, Barry Devereux*, Brian Murphy*

Queen's University Belfast
University Road, Belfast, Northern Ireland, BT7 1NN
{lwatson11, b.devereux, brian.murphy}@qub.ac.uk

Abstract

While there is a rich literature on the tracking of sentiment and emotion in text, modelling the emotional trajectory of longer narratives, such as literary texts, poses new challenges. This problem consists of two distinct elements – how much of the sentiment of a sentence in context is determined by the information contained in that sentence, and how much is determined by the previous history of sentences. Two key challenges in this work are identifying the features of the preceding context that are relevant to this sentence-by-sentence valence assignment task and determining the type of machine learning model which will perform best in predicting these sentence valences. As an initial exploration we apply a number of different feature sets using linear methods. We show that there is a small linear relationship between sentence valence history and the valence of the current sentence, which is statistically significant (P= 0.0001). These results show that we have captured a real effect in modelling the relationship between valence history and current sentence valence, the magnitude of explained variance is small. We intend to explore the application of models with non-linear capabilities in future work.

Keywords: sentiment analysis, fiction, narratives, valence, EDO

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 NLP, there is a rich literature on using lexical, semantic and structural information to infer an emotional tag for sentences and short passages. However modelling the emotional trajectory of narratives poses new challenges, in accounting for both the long distance effects of previous discourse, and the contextually subtle ways in which the high-level information conveyed by a text can influence a reader's emotional state. Our future aim is to develop a system that uses machine learning methods to automatically predict the emotional trajectory of literary texts such as novels. In order to understand what we mean by the phrase 'emotional trajectory' it is necessary to consider how the experience of emotion relates to the process of reading – what kinds of emotion people experience when reading and whether any given story can be said to have or relate a particular experience of emotion. Mar et al. (2011) highlight a number of ways in which our experience of emotion interacts with the process of reading – how moods affect our choice of reading material, the different ways we experience emotion while reading, and the lingering after-effects of emotions invoked when reading and their potential to affect our lives. Mar et al. propose a taxonomy of five narrative emotions experienced while reading: *emotions of sympathy*, *emotions of identification*, *emotions of empathy*, *relived emotions*, and *remembered emotions*. Of these five categories, the first three are most relevant to our study as they relate to emotions experienced in direct response to narrative events either in sympathy with the characters or through imagining ourselves in the characters' place. These three categories of emotions can be regarded as being properties of the story itself and, as such, are in some way contained in segments of the actual text, and therefore should in principle be retrievable from the text using quantitative methods (as opposed to the final two that relate to prior individual experience).

The field of sentiment analysis has begun to answer the question of how we can measure emotions in text, particularly with regard to commercial domains and social media. For example, work on sentiment analysis (i.e. the task of "automatically determining valence, emotions, and other affectual states from text."; Mohammed, 2016) has focused on product or movie reviews (Liu, 2010; Mohammed, 2016; Socher et al., 2013; Tai et al. 2015) or on the analysis of twitter feeds (Liu, 2010). Recent work using deep learning, and in particular recurrent neural networks (RNN) such as Long Short-Term Memory (LSTM) neural networks 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), they show particular promise for modelling high-level properties of natural discourse, such as literary texts.

However, most of the work on sentiment analysis makes use of large, readily available corpora of labelled data, which contain explicit rating values that can be used to train a classifier or neural network model (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). In the case of our study, no large dataset of pre-annotated literary text exists and so we start by developing a method which can learn to predict the emotional content at a given point in a story given only the preceding context and existing lexical resources (such as a hand-tailored sentiment dictionary, or a corpus-derived word-embedding). 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 as consisting of two distinct elements – how much of the sentiment of a sentence in context is determined by the information contained in that sentence (what the reader is reading now) and how much of the sentiment of a sentence in context is determined by the previous history of sentences (what the reader has been reading up to now). Drawing on Russell's two dimensional model of emotion (Russell, 1980 cited in Mohammed, 2015), consisting of valence (i.e., a positive-to-negative scale of emotional affect) and arousal (i.e. the intensity of the emotional response), we focus on modelling valence at the sentence level. We also explore the effects of incorporating semantic information through the inclusion of word embeddings.

To our knowledge, very little previous work has directly examined the influence of sentence history on the current sentence 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 (2009) takes a mean of all word valence values as representative of the valence value for different chunks of text e.g. sentence, paragraph, chapter, etc.

In this initial work we limit ourselves to modelling sentence level valence based on a single lexicon of sentiment (Whissell, 2009). We have chosen sentence level sentiment as the best unit of measurement for comparison with human annotations. Sentence-level valence is estimated as the mean of the sentence's word valences, and is modelled using linear methods, in terms of the mean valences of sentences in preceding context. 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, 2009; Whissel, 2003; Bestgen, 1990) and offers a computationally inexpensive way to begin this exploratory work.

2. Methodology

We train models that, at each sentence in the text, assign a measure of the emotional polarity at that sentence given information available in the preceding context. In order to establish a baseline for comparison we made some simple, explorations of the data at the level of individual words using existing lexicon resources describing word valence. We processed these lexicon-based valence ratings in two ways. First, we took a simple mean of the word ratings for each sentence as an overall measure for that sentence's valence rating. This basic measure showed a correlation (r^2) of 0.004 with the results of the human annotations (see Section 2.1.4). Next, we took a triangular weighted average over a backward-looking word window (consisting of a specified number of words) and took the valence and arousal values for each word in an end-of-sentence position to represent the sentence rating. We then correlated these modelled sentence ratings with human annotations. to determine the statistical relationship.

Our aim in this study is to improve upon these initial baseline results using data-driven methods. The key challenges are in identifying the features of the preceding context that are relevant to this sentence-by-sentence valence assignment task and in determining the type of machine learning model which performs best in predicting these sentence valences. Thus, our methodology proceeds in two parts. Firstly, we explore a number of different feature sets, to determine which kinds of information are most important for determining sentence-level valence. As well as information available in the sentence itself, we explore the scope of context relevant to inferring sentence valence. A given literary text is organised into subgroups of coherent parts in the form of chapters, paragraphs, and sentences, and within each of these subcategories the basic units will have a certain coherence of meaning, style, theme and sentiment. When a person reads a work of literature the emotional arc of the text builds in the form of dramatic tension and the emotion that the reader will be feeling at any given point in the story can be informed by the entire history of the text to that point in the form of chapter, paragraph and sentence content. Capturing the semantic and thematic information of this history may add a further level of depth to the model. This first stage of our study therefore focuses on the exploration of three different feature sets: (1) a history of sentence valence scores only (over a number of history window sizes), (2) a history of sentence valence combined with semantic information (i.e. semantic word embeddings, again over a number of history window sizes), and (3) a history of sentence, paragraph and chapter valence (where paragraph and chapter valence are calculated at the paragraph- and chapter level respectively, rather than at the sentence level, in order to keep the number of features relatively consistent across models). We fit a simple linear regression model using these three different feature sets to determine which features prove most informative in predicting sentence valence from history.

In the second part of the study, we use the best performing feature set from the first stage of the study to explore in further detail how the features relate to valence information. To this end, we compare the results using a simple linear regression model to a support vector regressor (in order to assess whether a model which is more robust to outliers can improve our predictions). We evaluate the emotional polarities predicted by the model by comparing the results to their corresponding target values. We used nltk to preprocess the texts and scikit-learn to implement the machine learning models.

2.1. Data and resources

2.1.1. Text for Human Annotation

The target text we selected for this preliminary study is *Harry Potter and the Philosopher's Stone: Chapter 9, The Midnight Duel* (Rowling, 1997). This text was chosen as it represents a piece of popular literary fiction for children which will be well known and accessible to most people and which most people would find emotionally engaging, but which at the same time still encompasses the complexity and linguistic variety of natural language (Wehbe et al., 2014).

2.1.2. Training Corpus

Project Gutenberg (https://www.gutenberg.org/) provides access to thousands of public domain books (copyright expired) in plain-text format. We selected a set of 100 books to use as training data. These books were chosen based on their similarity to the target text (works of literary fiction, many of them children's fiction). They shared with the target text 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.

2.1.3. Lexicons and lexical embeddings

In training our models, we used information about the emotional content of previous words in the sentences derived from Whissell's Dictionary of Affect in Language (the Revised DAL; Whissell, 2009). There are a number of other affect lexicons in popular usage in the field of sentiment analysis such as the NRC Emotion Lexicon (Mohammed and Turney, 2013), The Opinion Lexicon (Liu, et al., 2005), and AFINN (Nielsen, 2011), but these focus on a relatively limited lists of words thought to have a high emotional content. Approaches based on highly emotive words were initially adopted in psychology and are still predominant in sentiment analysis but Whissell (2009) showed that lexicons designed in this way did not achieve good enough coverage levels of target texts in natural language. The Revised DAL was created to target natural language and has been demonstrated to give 90% coverage of target texts in English (Whissell, 2009). While our initial explorations of our target text don't quite reach that level of coverage (59%), Whissell's DAL lexicon still has significantly better coverage than other lexicons (4% for each of the other lexicons mentioned above).

We generated sentence-by-sentence valence ratings for our target texts using data from the Whissell lexicon, where the valence for each sentence is estimated as an average over the valences for the constituent word in the sentence that are found in the lexicon. We then took these sentence-level valence ratings as the target values we hoped to predict for the current sentence's valence, using the history of previous sentence valence values to train our model. As described above, we also investigated the effect of using additional semantic features in our model, combined with the word valence ratings taken from the affect lexicon. These semantic features were derived from the GloVe word representation vectors available from the Stanford NLP group (Pennington, et al., 2014). Including this data allowed us to investigate how the semantic as well as affective properties of previous words influenced the text's evolving emotional trajectory.

2.1.4. Human Survey Data

We gathered human annotations for the target text using the online survey tool LimeSurvey (https://www.limesurvey.org). We asked nine adult native English speakers (six males and three females) to rate each individual sentence in the text for both valence and arousal. To preserve contextual information across sentences, the text was presented in paragraph groups with one paragraph per survey page. The paragraphs were further subdivided into sentences with each sentence representing a separate survey question. The survey consisted of 108 paragraphs (sequential in the novel), divided into 385 sentences. Below each sentence were two 5-point scales. The participants were advised to read the story naturally, taking the "side" of the protagonist (Harry). Where sentences may have seemed to contain both positive and negative emotions, or varying degrees of intensity, participants were instructed to rate them based on their feeling at the end of each sentence. Participants were also asked to complete a short, pre-survey familiarisation task where they were asked to rate a number of unconnected passages from other chapters of the same Harry Potter book, using the same rating scales. Once the data from the survey was collected a mean was taken of the participants' responses, and the mean human annotations were compared to the mean sentence scores derived from the Whissell lexicon showing a correlation (r^2) of 0.004.

2.2 Model training and evaluation

We took the Gutenberg texts and split them into training and test data. Our corpus consists 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. We then ran a linear regression over sentence history windows of 10, 50 and 100 sentences long respectively, in order to investigate the timescales at which previous information influences the valence of the current sentence.

To investigate the influence of lexico-semantic information on sentence valence, we then repeated these analyses including the wordspace vector information from the GloVe dataset as additional feature information. In these analyses, each sentence valence included in the history as predictors was represented by a 51 dimensional vector including the valence value for that sentence and the word vector information.

In order to investigate the effects of much longer scale contextual information, we next ran the regression based on a combination of sentence, paragraph and chapter valence history. To determine the appropriate window sizes for the sentence, paragraph and chapter history we created histograms to visualise the number of chapters, paragraphs and sentences in each book, chapter and paragraph respectively. Then we created a frequency distribution for each category and used this information to determine appropriate sizes for each window. We used a sentence window of 10 which had a 95% coverage of the sentences per paragraph in our dataset. We used a paragraph window of 50 which had a 73% coverage of the paragraphs per chapter in our data set. We used a chapter window of 50 which had an 81% coverage of the chapters per book in our data set.

The results of both the sentence window models and the sentence-paragraph-chapter (SPC) window models are found in Table 1 below.

Sentence Window	Model			
	Linear Regression		Linear Regression w/ semantic vector	
	Fit (r^2)	Prediction (r^2)	Fit (r^2)	Prediction (r^2)
10	0.0221	0.0210	0.0265	0.0168
50	0.0278	0.0259	0.0498	0.0106
100	**0.0292**	**0.0268**	0.0498	0.0067
SPC*	0.0284	0.0257		

*Sentence Paragraph Chapter History

Table 1: comparison of sentence history length, and of valence vs enriched lexical representation

We can see from this table that the basic feature set of sentence valence history only (over a window size of 100 sentences) proved the most informative in predicting current sentence valence. The addition of either word vector information or paragraph and chapter history information did nothing to improve the model. Our expectation was that the addition of semantic features would improve the model but our initial experiments have not supported this hypothesis. This may be due to the fact that the information necessary to predict current sentence valence is contained almost completely in the sentence valence history and that the information encoded in the semantic representation is redundant or irrelevant. However, it is also possible that the inclusion many additional parameters to our model has increased its tendency to overfit, which would explain why it has not generalised well to predicting on the test set. In future work we intend to address this problem by increasing the size of our training set, or through exploring the use of other models which are capable of learning from such a rich feature set over a relatively small amount of training data (e.g. by regularisation). Once we can establish a model that predicts well incorporating the semantic information it will be possible to investigate further which are the most pertinent contributory features. The beta values for the 100 sentence history window size are graphed in Figure 1.

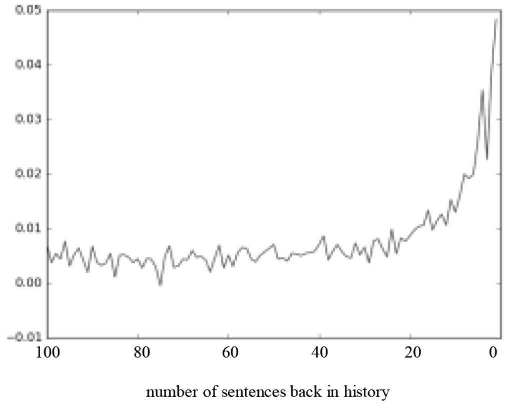

Figure 1: Regression beta values for 100 sentence history window

Taking the sentence valence history feature set, we then implemented a linear support vector regressor across the same range of sentence history window lengths. The results of this model, compared with the results of the linear regression, are compiled in Table 2.

Sentence Window	Model			
	Linear Regression		LinearSVR	
	Fit (r^2)	Prediction (r^2)	Fit (r^2)	Prediction (r^2)
10	0.0221	0.0210	0.0187	0.0210
50	0.0278	0.0259	0.0240	0.0257
100	**0.0292**	**0.0268**	0.0249	0.0262

Table 2: comparison of linear regression and linear support vector regression

From these results we can see that the simple linear regression model remains the best performing model. The linear SVR model, while also performing in the same range of values as the simple linear model, failed to improve upon it, indicating that the simple linear model is not overly sensitive to the presence of outliers.

3. Discussion and Future Work

We can see from these results that there is a small linear relationship between sentence valence history and the valence of the current sentence, which is statistically significant (P= 0.0001). While these results clearly show that we have captured a real effect in modelling the relationship between valence history and current sentence valence, the magnitude of explained variance is small. We believe that this may due to the limitations of a linear approach and the manual selection of features. We hope that a more complex model based on neural networks – capable of feature discovery, and of modelling non-linearities and complex relationships in the sentence valence history – will yield better results. We therefore intend to implement a number of neural network models such as Long Short-Term Memory Neural Networks (LSTM) (Hochereiter et al., 1997), a subclass of Recurrent Neural Nets, which can keep salient historical information in memory and make decisions about what information should be stored and/or replaced. Tai, Socher and Manning developed a version of LSTM implemented using a Tree structure which has proven to outperform other methods both on sentiment analysis and on predicting the semantic relatedness of two sentences (Tai, et al., 2015). We believe a similar approach will be better suited to analysing the complex processes involved in decoding the emotional content of a story in a way that is more representative of how narrative language works, for instance, by keeping relevant information about certain character or plot states in memory across passages or chapters.

4. References

Bestgen, Y. *Can Emotional Valence in Stories Be Determined From Words*. In: *Cognition and Emotion*, Vol. 8, no. 1, p. 21-36 (1994).

Finn Årup Nielsen. 'A new ANEW: Evaluation of a word list for sentiment analysis in microblogs', Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, 718 in CEUR Workshop Proceedings : 93-98. 2011 May. http://arxiv.org/abs/1103.2903.

Hochreiter, Sepp and Ju¨rgen Schmidhuber. 1997. 'Long Short-Term Memory'. *Neural Computation* 9(8):1735–1780.

Hsu CT, Jacobs AM, Citron FM, Conrad M. 'The emotion potential of words and passages in reading Harry Potter – an fMRI study'. Brain Lang. 2015 Mar; 142:96 – 114. doi: 10.1016/j.bandl.2015.01.011. Epub 2015 Feb 11.

Jockers, M. (2015) 'Revealing Sentiment and Plot Arcs with the Syuzhet Package'. http://www.matthewjockers.net/2015/02/02/syuzhet/

Liu, B. 'Sentiment Analysis and Subjectivity' 'Handbook of Natural Language Processing', Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010

Liu, B., Hu, M., and Cheng, J. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

Mar, R.A., Oatley, K., Djikic, M., Mullin, J. 'Emotion and Narrative Fiction: Interactive influences before, during and after reading', *Cognition & Emotion*, 25:5, 818-883, 2011.

Mohammed, S M., 'Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text',Emotion Measurement, Elsevier, ed: Meiselman, H., 2016.

Mohammad, S.M. Kiritchenko, S. and Zhu, X. 'NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets', In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), June 2013, Atlanta, USA.

Mohammad, S.M. 'From Once Upon a Time to Happily Ever After: Tracking Emotions in Mail and Books', Decision Support Systems, Volume 53, Issue 4, November 2012, Pages 730–741.

Mohammed, Saif. and Turney, P. 'Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon', *Computational Intelligence,* 29 (3), 436-465, 2013.

Pennington, J., Socher, R., and Manning, C., 2014. GloVe: Global Vectors for Word Representation., Empirical Methods in Natural Language Processing (EMNLP), 1532--1543

Rowling, J. K. *Harry Potter and the Philosopher's Stone*. London: Bloomsbury Pub, 1997.

Socher, R. Perelygin, A. Wu, J. Chuang, J. Manning, C. Ng, A. Potts, C. 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,' 2013, Conference EMNLP

Tai, K.S. Socher, R. and Manning, C. 'Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks', CoRR, abs/1503.00075, 2015. http://arxiv.org/abs/1503.00075

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, et al. (2014) 'Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses'. PLoS ONE 9(11): e112575. doi:10.1371/journal.pone. 0112575

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: Volume 96, Issue , pp. 990-992.

Whissell, C. 'Whissell's Dictionary of Affect in Language Technical Manual and Users Guide', *www.cs.columbia.edu/.*