Cognition-Cognizant Sentiment Analysis with Multitask Subjectivity Summarization based on Annotators' Gaze Behavior

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, Kuntal Dey
IBM Research AI, India
{abhijimi,srikanth.tamilselvam,riddasgu,senagar3,kuntadey}@in.ibm.com

Abstract

For document level sentiment analysis (SA), Subjectivity Extraction, *ie*., extracting the relevant subjective portions of the text that cover the overall sentiment expressed in the document, is an important step. Subjectivity Extraction, however, is a hard problem for systems, as it demands a great deal of world knowledge and reasoning. Humans, on the other hand, are good at extracting relevant subjective summaries from an opinionated document (say, a movie review), while inferring the sentiment expressed in it. This capability is manifested in their eye-movement behavior while reading: words pertaining to the subjective summary of the text attract a lot more attention in the form of gaze-fixations and/or saccadic patterns.

We propose a multi-task deep neural framework for document level sentiment analysis that learns to predict the overall sentiment expressed in the given input document, by simultaneously learning to predict human gaze behavior and auxiliary linguistic tasks like part-of-speech and syntactic properties of words in the document. For this, a multi-task learning algorithm based on multi-layer shared LSTM augmented with task specific classifiers is proposed. With this composite multi-task network, we obtain performance competitive with or better than state-of-the-art approaches in SA. Moreover, the availability of gaze predictions as an auxiliary output helps interpret the system better; for instance, gaze predictions reveal that the system indeed performs subjectivity extraction better, which accounts for improvement in document level sentiment analysis performance.

1 Introduction

Document level Sentiment Analysis (SA) has been a well-studied problem. It deals with predicting the polarity of the opinion expressed by a user/reviewer in the form a discourse, typically spanning over a few paragraphs. In this work, we propose yet another solution to document level SA; specifically, we tackle binary sentiment classification, for instance, classifying a movie review as positive ("thumbs up") or negative ("thumbs down"). Though sentiment expressed in an opinion can be different for different aspects (*e.g.,* "quality of music" in a movie review), and aspect-based SA is a separate branch of research, SA systems are often required to predict the sentiment polarity based on one central aspect that the overall opinion revolves around. Such binary SA systems have applications galore in social media analysis, recommender systems, e-commerce settings, to name a few.

The central challenge for such SA systems is to figure out the portions of the input review that are *subjective*, *i.e.,* contain opinion and not facts, and are *relevant i.e.,* express opinion regarding the central aspect. This process is called *Subjectivity Extraction* (Pang and Lee 2004). A subjective extract of a movie review, per se, would not contain facts (*e.g., Jim Caviezel plays Dantes*), or opinions that are less connected to the central aspect (*e.g., "Jim Caviezel was excellent as the count, and you loved to hate Guy Pearce as his friend-gone-enemy"*). Subjectivity Extraction remains a problem hard to solve, as it demands tremendous amount of world knowledge and reasoning. A few works, such as Pang and Lee (2004) and (Mukherjee and Bhattacharyya 2012), successfully extract facts from opinion before performing SA, but extracting the relevant subjective portions still remains a difficult problem.

Unlike machines, humans are quite apt in extracting subjective summaries from reviews. This is evident from their reading behavior, while they read and infer sentiment expressed in an opinion. Mishra, Joshi, and Bhattacharyya (2014), through a pilot study, observe that, human annotators perform subjectivity extraction in such a way that the extracts are gradually formed and revised during reading of the document. This is supported by the eye-movement behavior of the annotators. Eye-movement attributes such as gaze-fixations and regressive saccades, provide evidences regarding what portions of the text needs to be retained as subjective extracts and how subjective extracts gradually build up before decision regarding sentiment is finalized. In summary, document level sentiment analysis, for humans, is an interplay between multiple tasks such as subjectivity extraction and sentiment inferencing.

Can machines learn to read like humans, and perform sentiment inferencing? We attempt to answer this question by proposing a multitask recurrent neural network architecture that learns to predict sentiment by simultaneously learning to predict gaze. Additionally, the system also learns to perform an auxiliary linguistic tasks of part-of-speech (POS) tag prediction. The linguistic task is included to make the system aware of the lexical and syntactic properties of the text. Our architecture is based on shared layers of Long

Short Term Memory (LSTMs) (Hochreiter and Schmidhuber 1997), built on top of word embeddings - the task specific classifiers are branched out of the shared layer. The gaze-prediction task and auxiliary linguistic task happen at a word level (analogous to the problem of sequence labeling) and the sentiment prediction happens at a document level. With this composite multi-task network, we obtain performance competitive with or better than state-of-the-art approaches in SA. Moreover, availability of gaze prediction as an auxiliary output helps interpret systems better; for instance, gaze predictions reveal that the system indeed performs subjectivity extraction better, impacting the document level sentiment decision, and hence, the accuracy.

The contributions of this paper are the following.

- We present a novel multi-task learning approach for document level sentiment analysis, that considers the modality of human cognition along with text, making it a human-inspired AI system.
- Our method is supervised, but is designed to leverage multiple disjoint datasets available particularly for individual tasks. This eliminates the requirement of multi-labeled datasets, which is often not available.
- Eye gaze data, which is expensive, is only required during training the system, and not during testing.
- While techniques like Dropout and Regularization have been used to address overfitting, our system, by design, naturally avoids overfitting, due to the inclusion of multiple tasks.

2 Related Work

Text sentiment analysis (SA) has been a long-standing area of research, for short text level as well as document level sentiment analysis. In an early work, Hatzivassiloglou and McKeown (1997) identified the sentiment polarity orientation of adjectives, using conjunction constraints, in a four-step supervised learning algorithm, and applied on Wall Street Journal corpus. Several works have been proposed towards SA of user-generated text, *viz.,* movie reviews (*e.g.* IMDB[1], Rotten Tomatoes[2]), such as Pang and Lee (2004), Whitelaw, Garg, and Argamon (2005), Denecke (2008), Maas et al. (2011) and Palkar et al. (2016), and online social network content (*e.g.* Twitter[3]), such as Wilson et al. (2005), Agarwal et al. (2011), Barbosa and Feng (2010), Kouloumpis, Wilson, and Moore (2011), Khan, Atique, and Thakare (2015), Le and Nguyen (2015) and Zimbra, Ghiassi, and Lee (2016). Lin and He (2009) propose an LDA (Blei, Ng, and Jordan 2003) based topic-model to jointly detect topic and sentiment from text, and apply on movie review datasets.

Recent works, such as Kim (2014), Socher et al. (2013) and Hassan and Mahmood (2017), have applied deep learning frameworks, including convolutional neural networks (CNN) and embodiments of recursive neural networks (RNN) such as long short-term memory networks (LSTM), towards the task of SA. Socher et al. (2013) create a tensor-based RNN network to improve the classification accuracy of single-sentence short documents, and propose a sentiment treebank in the process. And in a CNN-based approach, Kim (2014) perform a simple convolution with multiple filter widths, and features obtained via the convolution operations on shorter text.

Works, such as Mishra et al. (2016) and Mishra, Dey, and Bhattacharyya (2017a), have taken the eye gaze of readers into the account, for the tasks of sentiment classification and sarcasm detection in sentiment text respectively. These works record the eye-movements of readers as they read through text. They perform supervised learning from text attributes and gaze attributes such as fixations and saccades, and graph attributes obtained by analyzing the scanpath of readers gaze movements in the reading session. A recent work by Mishra, Dey, and Bhattacharyya (2017b) further investigated the effectiveness of deep CNNs for sentiment and sarcasm analysis. In this work, the authors convolve the semantic embeddings of text presented to the readers in a single dimension, and convolve the eye movement patterns observed across the different participants for reading a given piece of text in two dimensions, and finally combine the convolutions to obtain the overall text classification. A detailed literature survey for SA, has recently been conducted by Yadollahi, Shahraki, and Zaiane (2017).

Extraction of subjective text segments have been used for document-level SA. Extracting subjective text segments poses a tremendous challenge, that only a few works have attempted. Pang and Lee (2004) uses this approach. They apply a graph-mincut based technique to separate the subjective portion of the text from the irrelevant objective portions. Mukherjee and Bhattacharyya (2012) show that, for sentiment prediction of movie reviews, subjectivity extraction may be used to discard the sentences describing movie plots, since they do not contribute towards the speaker's view of the movie. While these techniques mostly rely on removing non-subjective portions of text to identify good subjective extracts, they may not discard subjective sentences unrelated to the overall document sentiment.

Our work is inspired by an eye-tracking based pilot-study on Subjectivity Extraction (Mishra, Joshi, and Bhattacharyya 2014), which shows that, when annotating documents for sentiment, humans employ subjectivity extraction in two forms: *anticipation* and *homing*. The choice between anticipation and homing depends on the way sentiment changes in these documents. In *linear* documents, where all or most sentences are of the same polarity, readers tend to not read certain portions of the document at all. They read sentences of the same polarity appearing in the same review previously, and anticipate that to continue. For such documents, the authors perform *subjectivity extraction through anticipation*. For *oscillating* documents, where sentiment changes through most sentences, the reader first reads an entire document, and then re-visits a subset of (semantically overlapping) sentences in the document before arriving at a judgment about the sentiment of the document. For these, the authors perform *subjectivity extraction through*

[1] http://www.imdb.com

[2] https://www.rottentomatoes.com

[3] https://twitter.com

homing.

To empower our system to learn patterns related to subjectivity extraction from gaze data, we adopt a multi-task LSTM, to jointly perform the main task of SA along with auxiliary tasks of part-of-speech (PoS) tagging. This is akin to the multi-task LSTM based approach of Klerke, Goldberg, and Søgaard (2016) which also learns to predict eye gaze with respect to the main task of sentence compression.

3 Subjectivity Extraction and Eye-movement

As discussed in the earlier sections, extracting appropriate subjective extracts from a given document is a key step to SA. In this regard, it is important to lay foundations for what should ideally constitute the subjective extract of a document. As per our observations in the dataset we use, an opinionated document may contain four components:

1. Relevant subjective sentences: Sentences that contain opinion about the aspect on which sentiment analysis has to be performed.
2. Irrelevant subjective sentences: Contain opinions about other aspects and are not related to the aspect on which sentiment analysis has to be performed.
3. Irrelevant objective sentences: These are objective sentences and they contain factual details but not opinion.
4. Anchor Sentences: Sentence that contain complete information about the overall polarity of the text. Anchor sentences are a subset of Relevant subjective sentences.

An opinion will necessarily contain (1), whereas (2), (3) and (4) might or might not be present. To explain this in detail let us consider the following example review with respect to the aspect "movie":

> *Saw this movie on the 28th of December. I walked out of the theater very, very, very satisfied with the movie. The audience was the worst audience I've ever sat through a movie with it. If the audience is bad, it can ruin the movie, and make you like it half as much. That's probably why it's only my second favorite movie. (My favorite being Office Space) Though this movie is rated R, it really isn't that bad. There is blood, but no gore. When someone gets stabbed, naturally, they're going to bleed. When someone gets shot, naturally, they're going to bleed. But, they're flesh isn't naturally going to be split apart. This movie keeps it realistic. To tell you what it's about: it's about an ex-civil war captain. He goes to Japan to teach the Japanese soldiers American tactics. In their first battle they aren't ready and get defeated. The captain gets captured and taken to a place with many Samurai. At first, he's their enemy. He then learns the way of the Samurai, and befriends the Samurai. To tell you anymore would be to ruin the movie. But I can tell you this much: go see this movie- you won't regret it.*

Here, different components of the text are represented in various colors. The blue colored sentences are objective in nature as they correspond to factual details (date of watching, plot summary *etc.*). Color "red" represents irrelevant subjective portions. As one would notice, the reviewer has a negative opinion about the audience in the theater which does not affect the overall sentiment of the person towards the movie. The portions in "black" and "green" represent the subjective extract of the text with the green colored sentence being the anchor one.

Ideally, when a person reads a text like above with a goal of understanding the underlying sentiment, the maximum information about the opinion should come from the anchor sentence followed by other relevant subjective portions, portions that are part of the subjective extract. The irrelevant portion should draw least amount of attention. We verify these hypotheses through our eye-tracking studies as explained below.

3.1 Creation of a Document Level Sentiment Analysis and Eye-movement Dataset

Our dataset consists of 23 movie reviews from Amazon movie review dataset[4]. The reviews were selected as per ratings given by the writer. There were 12 reviews with a rating of 5, 9 reviews with a rating of 1 and 1 review each with a rating of 4 and 3. The average number of sentences per review was 11.65 (standard deviation of 4.37) with an average of 22.03 words per sentence (standard deviation of 7.56). We obtained annotation from 33 human participants. To ensure language proficiency, these participants were chosen such that one of these criteria was met: (a) The participant had completed an English language proficiency examination among IELTS, TOEFL, etc., or (b) The participant had completed a graduate degree with English as the primary language of instruction. The mean age of our participants was 25 years with a standard deviation of 2 years.

The task assigned to the participants was to read one document at a time and annotate it with the appropriate sentiment label. While they read the document, their eye-movement behavior was recorded using an SR-Research Eyelink 1000 Plus eye-tracker, with a sampling rate of 500 hz. The experiment was controlled to minimize noise in eye-movement recording and fatigue associated with reading a large number of paragraphs in one sitting. Participants had to undergo a "practice sentiment reading" session to get used to the task. Details regarding the experiment can be found out in the supplementary material.

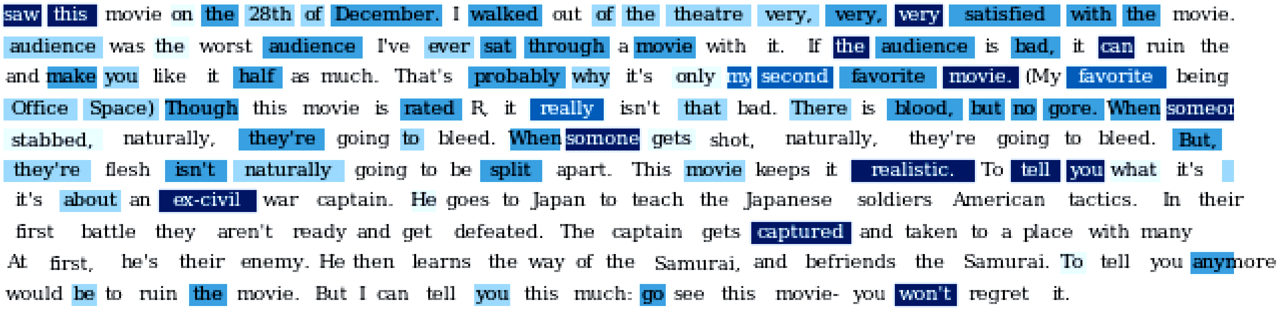
The annotation statistics are as follows: 13 out of 23 documents have 100% annotations[5]. 8 participants annotate all documents correctly, while 16 participants get one annotation wrong. 8 participants annotate two documents incorrectly.

3.2 Analysis of Fixations

We use the word-level *first fixation duration* (FFD) profile of the participants. Word-level FFD is defined as the time period that an individual looks at the given word, when their eye gaze fixates for the first time on the word in their reading

[4] https://snap.stanford.edu/data/web-Movies.html

[5] for the rest of the documents eye-movement data from some of the participants was not fit for use due to unusual shifting of gaze due to head movement



saw this movie on the 28th of December. I walked out of the theatre very, very, very satisfied with the movie.
audience was the worst audience I've ever sat through a movie with it. If the audience is bad, it can ruin the
and make you like it half as much. That's probably why it's only my second favorite movie. (My favorite being
Office Space) Though this movie is rated R, it really isn't that bad. There is blood, but no gore. When someor
stabbed, naturally, they're going to bleed. When somone gets shot, naturally, they're going to bleed. But,
they're flesh isn't naturally going to be split apart. This movie keeps it realistic. To tell you what it's
it's about an ex-civil war captain. He goes to Japan to teach the Japanese soldiers American tactics. In their
first battle they aren't ready and get defeated. The captain gets captured and taken to a place with many
At first, he's their enemy. He then learns the way of the Samurai, and befriends the Samurai. To tell you anymore
would be to ruin the movie. But I can tell you this much: go see this movie- you won't regret it.

Figure 1: Gaze-Word heatmap based on normalized first fixation duration (FFD) of a participant for a movie review. Color intensity if positively correlated with amount of fixation duration spent on the area of interest (words)

process. Following Klerke, Goldberg, and Søgaard (2016), FFDs are first discretized into six bins as follows.

Label = 0, if $FFD = 0$
Label = 1, if $FFD < \mu - \sigma$
Label = 2, if $FFD < \mu - 0.5\sigma$
Label = 3, if $FFD < \mu + 0.5\sigma$
Label = 4, if $FFD > \mu + 0.5\sigma$
Label = 5, if $FFD > \mu + \sigma$

Here μ and σ are mean FFD and standard deviation for a particular reader, and for a particular document. Note that only non-zero values contribute to the calculation of mean the standard deviation (SD). After transforming the absolute FFDs into discretized FFD labels, we plot the heatmap profile of FFDs on the text.

As seen in Figure 1, which shows the FFD heatmap profile of one participant, a significant portion of the words are skipped during the reading process (white in color), while some words are fixated upon for long (dark blue) and the others for lesser (light blue). This is a general trend that we observe in sentiment oriented reading, where, mining clues related to the overall sentiment of the document proved to be sufficient. This results in less number of of fixations as opposed to what is seen in comprehension oriented reading related datasets (*e.g.,* The Dundee Corpus (Kennedy, Hill, and Pynte 2003)). Moreover, for most of the documents, The first-pass reading turns out to be decisive and FFDs on relevant subjective and anchor sentences turn out to be higher. So FFD seems to be an important gaze-based aspect to consider for the task. On the other hand, we empirically observe that repeat fixations are towards words that are mostly sentiment-neutral (objective), and thus are irrelevant for the task of sentiment analysis. Regressive saccades, that form the basis of regressive (repeat) fixations, are thus also not significant with respect to the task at hand. Thus, we limit the eye movement features to FFD, and don't consider repeat fixations and saccadic attributes. Note that, this makes document level SA a task different from sentence and short-text level SA, where the prior works in the literature, such as (Mishra et al. 2016), (Mishra, Dey, and Bhattacharyya 2017a) and (Mishra, Dey, and Bhattacharyya 2017b), have found such repeat fixations as well as saccadic movements (including regressive saccades) to be effective.

4 Neural Architecture for Multitask Learning

We design a multi-layer Recurrent Neural Multi-task architecture with document level sentiment polarity prediction as the main task and gaze and part-of-speech prediction as auxiliary tasks. Figure 2 depicts the design. Words (in the form of *one-hot* representation) in the input text are first replaced by their embeddings of dimension K (i^{th} word in the sentence represented by an embedding vector $x_i \in \mathbb{R}^K$). To tackle length variations, padding is applied wherever necessary, thereby transforming the sequence to a fixed length tensor (say N).

The embeddings are then passed to a couple of layers of bidirectional LSTMs (BiLSTM). BiLSTMs have already proven to be useful in capturing context better than the unidirectional variant, and have been used for fine grained sentiment analysis (Liu, Joty, and Meng 2015). For each timestep that corresponds to an input word, the word along with the contextual information coming from the LSTM cells from both the directions, from the previous time steps are encoded together, to form a vector $v_i \in \mathbb{R}^V$) of V dimensions.

$$v_i = BiLSTM(x_i, layers = 2), i \in N \tag{1}$$

The function BiLSTM can be expanded to a series of differentiable equations, as given by Hochreiter and Schmidhuber (1997).

Now, the multi-task setup allows us to fork multiple branches pertaining to the main and auxiliary tasks. The description for each task goes as follows.

1. **Document Sentiment Polarity Prediction (Main Task):** For this task, the forked branch takes all the encoded vectors obtained across all N timesteps and predicts sentiment labels (-1 or +1) as output. It first extracts the most significant components out of the encoded vectors by performing a *max-pooling over time* operation (Collobert et al. 2011). This will result in a fixed length vector c of size N.

$$c = [c_1, c_2, c_3, ..., c_N] \tag{2}$$

The features are then passed to a `softmax` layer, the out-

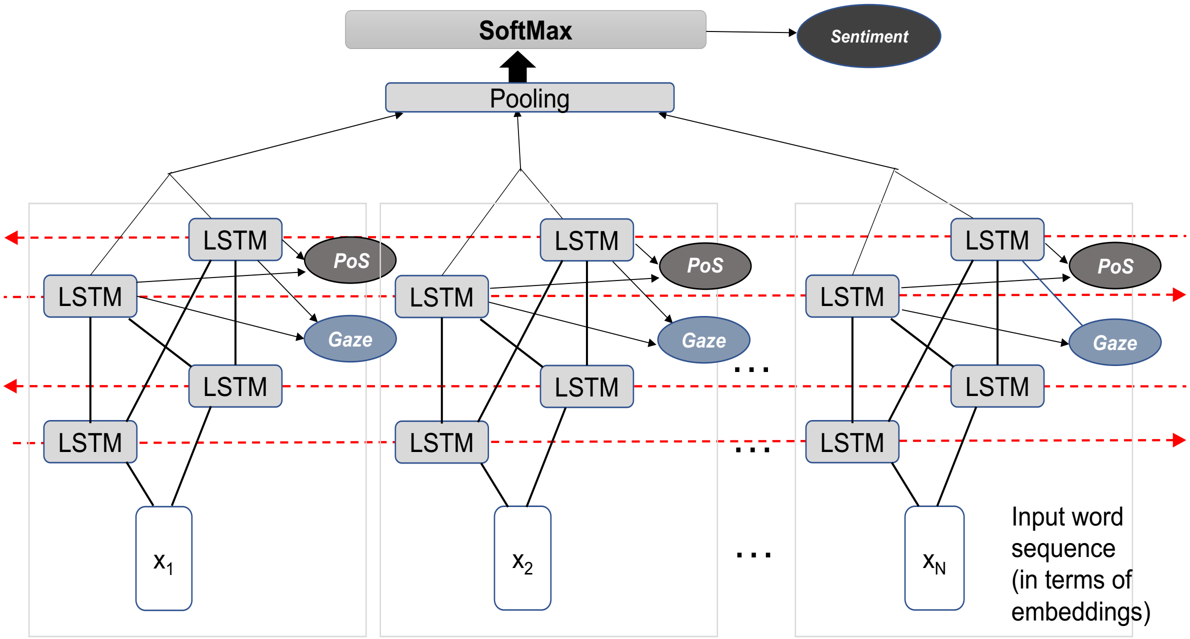


Figure 2: Deep multitask model with Sentiment Analysis as primary task and Gaze and Linguistic artifact prediction as secondary task.

put of which is the sentiment label y.

$$\begin{aligned}\hat{y} &= \arg\max_{y} P(y = j|c) \\ &= \frac{e^{c^T w_j}}{\sum_{k=1}^{|J|} e^{c^T w_k}}, J \in \{-1, 1\}\end{aligned} \tag{3}$$

During training, the optimizer optimizes the *binary cross entropy* loss between the predicted and the observed values of y.

2. **Gaze Prediction (Auxiliary Task1):** Gaze prediction in our setting is a *sequence labeling* task, where the First Fixation Duration (FFD) on each word is predicted. For this the encoded vector (analogous to eq. 1) for each time step is passed to a SOFTMAX layer (eq. 3). Gaze labels can take values between $[1-5]$, as discussed in Section 3). For loss computation, *categorical cross entropy* is used.

3. **Part of Speech (PoS) Prediction (Auxiliary Task2):** Certain parts of speech (such as adjectives and adverbs) have proven to be more informative about the sentiment (Benamara et al. 2007) of the text. Our model is made aware of this through the introduction of PoS tagging as one of the auxiliary tasks. Like gaze prediction, PoS tagging is also considered as a sequence labeling task, and the forked component of the network is of same nature (with a similar loss function) as the one used for gaze prediction.

For each step in the training process, a task out of the three possible tasks is chosen at random, followed by a selection of a random batch of data for the task. The forked model specific to the task predicts the label, suffers a loss with respect to the true labels, and the model parameters are updated. When the task changes, parameters of the shared layers are copied and set as the initial parameters for the forked model of the new task. Embedding layers are initialized using pre-trained embeddings and are not updated through out the training.

It is worth noting that, the tasks (especially the auxiliary ones) can be “composite” in nature. For example, for auxiliary task 2, one could consider joint prediction of PoS and syntactic and semantic properties such as dependency tags, semantic roles *etc.*, which should intuitively help in tackling linguistic nuances better. Similarly, gaze prediction can be a joint prediction of fixation durations and other gaze attributes like saccade amplitude, direction *etc.*, which are considered to be important in “reading” research (Rayner 1998). However, in our setting, we observe that inclusion of multiple and composite tasks results in over-generalization, thereby decreasing the model performance. Hence, we stick to only the basic tasks.

5 Experiment Setup

We now share the details of our experiments below.

5.1 Dataset

For evaluating our multitask model variants, we rely on the Movie Review dataset of 2000 documents, released by Pang and Lee (2004). This dataset is not only a benchmark one for comparing against various algorithms, it is also suitable for comparing how the gaze based subjectivity extraction (discussed in Section 6) compares against the extract obtained

Model	PL2000	IMDB25K
Mincut (Pang and Lee 2004)	87.15	N/A
Embeddings (Maas et al. 2011)	88.90	88.89
Our Variants		
Only Sentiment (unitask)	88.037	89.431
Sentiment + PoS	87.686	89.44
Sentiment+PoS+Gaze	89.014	89.428

Table 1: Results (in terms of accuracy) for different model configurations for two test datasets (PL2000 by Pang and Lee (2004) and IMDB25K by Maas et al. (2011)) . *Sentiment*→ Document level sentiment prediction task, *PoS*→ PoS prediction task, and *Gaze*→ Prediction of FFD.

using the initial graph min-cut based method proposed by Pang and Lee (2004). Though Pang and Lee (2004) used a 10-fold cross validation in their setting, we believe considering only 2000 instances for cross validation in a neural setting like ours will leave our model with less amount of data and significantly more number of parameters to learn. This may result in over fitting and thus, reduction in test-accuracy[6]. This is exactly why, the performance of our system may not be directly comparable to some of the contemporary systems such as Kim (2014) and Johnson and Zhang (2015). We would like to remind our readers that the sole purpose of this study is to gain a first level insight into whether inclusion of a cognitive task of gaze prediction in a multitask setup helps in improving document level subjectivity realization better or not.

For training the model, we use the IMDB movie review dataset from Maas et al. (2011), and use 25K documents available for training. Moreover, we also evaluate the performance of our models on 25K test data to compare with Maas et al. (2011). The gaze data, as discussed in Section 3.1, comes from a different source (*i.e.,* Amazon), and the reviews do not overlap with the dataset used for training or evaluation. This ensures that any improvement in performance observed by inclusion of gaze data is not because of any overlap across datasets. After removing noise in gaze data (unreasonable fixations and shifts that typically observed because of head movement during reading), the total amount of unique instances of Documents and FFD sequences turn out to be 757.

5.2 Hyperparameters

We use `word2vec` (Mikolov, Yih, and Zweig 2013) embedding with size set to 100. We use bidirectional LSTM with 2 layers and hidden size 100. Batch size was set to 128. The learning rate is reduced by a factor of 0.1 with a patience of 10 epochs.

5.3 Regularization

For regularization *dropout* is employed between the LSTM layers with a constraint on l_2-norms of the weight vectors (Hinton et al. 2012). Dropout prevents co-adaptation of hidden units by randomly dropping out - i.e., setting to zero - a proportion p of the hidden units during forward propagation. We set p to 0.3.

5.4 Training

We use ADAGRAD optimizer (Duchi, Hazan, and Singer 2011), with a learning rate of 0.1. The input batch size is set to 128 and number of training iterations (epochs) is set to 30. 10% of the training data is used for validation.

5.5 Pre-trained Embeddings

Initializing the embedding layer with pre-trained embeddings can be more effective than random initialization (Kim 2014; Liu, Joty, and Meng 2015). We combine all the textual portions of our datasets and used the data for pre-training. The embeddings are learned using skip-gram based `word2vec` learning mechanism (Mikolov, Yih, and Zweig 2013), with parameters `embedding_size`,`window_size` and `min_count` set to 100, 5 and 5 respectively.

Implementation of the multi-task model has been done using PyTorch[7] and for embedding learning we use the Gensim library (Řehůřek and Sojka 2010). For getting, PoS labels, we use the Spacy[8] PoS tagger (with Penn Tagset) which is fairly accurate.

We would like to bring it to the readers notice that many of our model hyper-parameters are fixed by trial and error and are possibly good enough to provide a first level insight into our system. Tuning of hyper-parameters might help in improving the performance of our framework, which is on our future research agenda.

6 Results and Discussion

Table 1 presents the results. It is clear that the addition of gaze as a prediction task helps in improving the performance of the system on PL2000 dataset (statistically significant)[9]. However, the uni-task model variant does well for the larger test dataset (IMDB25K). We observe that the PoS tagging accuracy of the system, in the multitask settings is often

[6] We observe a reduction of accuracy of around 15%, for the non-multitask baseline

[7] http://pytorch.org

[8] https://spacy.io/

[9] The difference of accuracy between the best and the second best system is statistically significant with $p < 0.05$, as confirmed by a two-tailed McNemar test

pretty low. This might be affecting the overall results for our multitask variants. One possible solution towards this would be to consider coarse-grained tags (Say just Noun, Verb, Adjective and Adverb) instead of the fine-grained tags that are used in the current experiment, which might be confusing the multitask labelers.

It is interesting to note that gaze task (even though trained on a small amount of data of just 757 instances) has a positive impact on the overall performance. Since, one of the main objective of including gaze prediction as a task was to interpret the model's behavior, we consider the gaze prediction output to prove / disprove hypotheses regarding whether the sentiment decision inferred by the model depends on how good the gaze predictions are on the input text (which, in turn, corresponds to how good / bad the subjective extracts are). A qualitative analysis on documents for which prediction of gaze based multitask model is better than only the gaze based model affirms that, most of the predicted long duration FFDs (Label>=4) are actually predicted on relevant subjective sentence. For the example mentioned in Section 3, the predicted long duration FFDs are on sentiment bearing words such as "satisfied", "favorite" appearing in the relevant subjective sentences and a few objective words such as "stabbed" and "shot" that implicitly convey negative sentiment. This kind of output interpretation helps us realize the impact of the gaze prediction task on the subjectivity extraction process.

7 Conclusion

In the current paper, we proposed a multilayer multitask LSTM for a main task of document-level sentiment analysis, and subtasks of word-level part-of-speech tagging and predicting eye gaze behavior of participants in form of first fixation durations. The framework undergoes simultaneous learning to enable the joint prediction, wherein the lower layers are architecturally shared across the tasks. In the training phase, it learns to predict human gaze given a document, attempts to closely follow the human cognition trajectory, thus inherently attempting to utilize capabilities of humans of extracting relevant subjective extracts from the document. The learning from the gaze prediction task is ingrained in the model, so that the framework can be used in real-life without collecting readers' eye gaze data during operations (testing phase). Further, predicting gaze as an auxiliary output enables us to interpret the system better; for instance, gaze predictions reveal that the system indeed performs subjectivity extraction better, which plays a key role in the document level sentiment decision. Our results with gaze data are competitive to well known methods for document level SA. However, the key takeaway from our experiments is that, linguistic tasks like Sentiment Analysis that deal with the pragmatic aspects of language processing can be tackled better in a multi-task setup, with the help of auxiliary cognitive tasks like gaze prediction. Our future plans include exploring more sophisticated multitask setups, such as adversarial multitask learning (Liu, Qiu, and Huang 2017) with a view to extracting useful task-invariant representations from both gaze and text data. Expanding gaze data for multitude of domains and performing cross-domain sentiment analysis using our multi-task setup is also on our agenda.

Acknowledgment

We thank the members of CFILT Lab, especially, Mr. Diptesh Kanojia IIT Bombay for their help and support for our experiments. Our special tanks to Dr. Pushpak Bhattacharyya, Professor, Department of CSE, IIT Bombay for his guidance.

References

Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passonneau, R. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, 30–38.

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 36–44.

Benamara, F.; Cesarano, C.; Picariello, A.; Recupero, D. R.; and Subrahmanian, V. S. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008.*, 507–512.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159.

Hassan, A., and Mahmood, A. 2017. Deep learning approach for sentiment analysis of short texts. In *Control, Automation and Robotics (ICCAR)*, 705–710. IEEE.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 174–181.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

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

Johnson, R., and Zhang, T. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 103–112.

Kennedy, A.; Hill, R.; and Pynte, J. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Khan, A. Z.; Atique, M.; and Thakare, V. 2015. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* 89.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Klerke, S.; Goldberg, Y.; and Søgaard, A. 2016. Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! *International AAAI Conference on Web and Social Media (ICWSM)* 11:538–541.

Le, B., and Nguyen, H. 2015. Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering*. 279–289.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384.

Liu, P.; Joty, S.; and Meng, H. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1433–1443.

Liu, P.; Qiu, X.; and Huang, X. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1–10.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 142–150.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Mishra, A.; Kanojia, D.; Nagar, S.; Dey, K.; and Bhattacharyya, P. 2016. Leveraging cognitive features for sentiment analysis. 156.

Mishra, A.; Dey, K.; and Bhattacharyya, P. 2017a. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 377–387.

Mishra, A.; Dey, K.; and Bhattacharyya, P. 2017b. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 377–387.

Mishra, A.; Joshi, A.; and Bhattacharyya, P. 2014. A cognitive study of subjectivity extraction in sentiment annotation. In *WASSA, ACL*, 142–146.

Mukherjee, S., and Bhattacharyya, P. 2012. Wikisent: Weakly supervised sentiment analysis through extractive summarization with wikipedia. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 774–793. Springer.

Palkar, R. K.; Gala, K. D.; Shah, M. M.; and Shah, J. N. 2016. Comparative evaluation of supervised learning algorithms for sentiment analysis of movie reviews. *International Journal of Computer Applications* 142.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124:372.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; 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*, 1631–1642.

Whitelaw, C.; Garg, N.; and Argamon, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 625–631.

Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, 34–35.

Yadollahi, A.; Shahraki, A. G.; and Zaiane, O. R. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50(2):25.

Zimbra, D.; Ghiassi, M.; and Lee, S. 2016. Brand-related twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1930–1938.