0.1 Targeted Sentiment Analysis

0.1.1 What is targeted sentiment analysis?

Targeted sentiment analysis is a fine-grained text-classification task which stems from the broader, more general, document, or sentence, level sentiment analysis. The former extends on the latter by taking into consideration a particular target or aspect within the context of the document, and aims to identify the sentiment with respect to this target or aspect [**?**], [**?**], [**?**].

It is often the case in the literature that when referring to a *target*, this would be a particular noun or subject within the phrase while an *aspect* can be a more general area or concept that the phrase touches on, without referencing in the literal sense. Consider a sentence such as, "The waiting times were long however the ravioli were simply to die for", a plausible *target* that could be considered is "waiting times" for which the statement conveys a negative sentiment. Alternatively, the phrase could be assessed with respect to an *aspect* such as "food quality", for which a positive sentiment is conveyed even though the precise term "food quality" is only implicitly implied.

It is evident that, separating itself from sentence-oriented sentiment analysis, target or aspect based sentiment analysis requires the careful consideration of the target or aspect in question along with its context. The extent of this fact was initially demonstrated by [**?**], whose work demonstrated that a staggering 40% of errors within the field of targeted-sentiment analysis could be attributed to the lack of consideration of the target or aspect [**?**].

0.1.2 What is the importance of targeted sentiment analysis?

Due to the proliferation of social media networks and online shopping, opinions voiced from users on specific topics, products, services and events have never been as readily available for data mining. The value in having the means to accurately gauge public interest and opinion of very specific topics of interest on such a phenomenal scale cannot be understated. From those in the public sector, such as electoral campaigns who seek to obtain a clearer picture of their constituents' strongest held opinions and expectations, to private businesses who wish to employ the most effective advertising campaign for their products and services, all of these objectives rely heavily on being as cognizant on public sentiment as possible. [**?**]

Over time the content of these online text sources has become more sophisticated and richer in information. Changes in social media platforms such as *twitter*'s decision to raise the character limit of tweets results in the same unit of data conveying up to twice the amount of information. As this availability increases, so to must the resolution at which this information is processed, so as to keep pace with the needs of both producers and consumers alike. This phenomenon further pushes

the need to focus on opinion mining at a finer-grained level, perfecting the ability to discern varying sentiments towards separate targets within the same phrase.

0.1.3 What are the challenges of targeted sentiment analysis?

As with any task that requires a deeper understanding of the intricacies of language, there are many challenges that face target-oriented sentiment analysis. Many of these challenges are inherent to parsing the structure of a language such as sarcasm, where sentences such as "Nice perfume. You must shower in it." [**?**] are composed entirely of positive-sentiment bearing words while expressing a negative sentiment. These notwithstanding, the informal nature of the majority of text data found on social media platforms (upon which this task frequently focuses), supplements these challenges with its own.

Colloquialisms and social short-hands are a commonplace within social media networks where many users intend on conveying as much information in as little characters as possible, particularly in situations where this number is capped. This phenomenon also leads to intentional, as well as unintentional, spelling errors which further obscures that data for any prospective machine learning model that does not account for these circumstances.

Along with these challenges, the literature also presents a number of obstacles and particularly problematic instances that need to be taken into account when approaching the task of targeted sentiment analysis. Comparative opinions are one such circumstance where the sentiment being conveyed is obscured by another subject. [**?**] report challenges of this sort, with phrases such as "I've had better Japanese food at a mall food court".

Other common challenges that are pointed out in [**?**], are negation and conditional situations, citing the example "but dinner here is never disappointing, even if the prices are a bit over the top", where the sentiment towards the target cannot be easily deduced from the various syntactic structures present.

Moreover, the particular case of expressions which consist of multiple words needs to be given special care. Various approaches that employ word embeddings operate on the word as the atomic unit of operation, and would therefore struggle to correctly model an expression such as "die for" in "the ice cream to die for" [**?**] from its constituents. [**?**] also also stress the significance of this issue, and argue that it has not been given sufficient attention, particularly when modelling *targets* that also consist of multiple words.

When considering the opinion of a sentence towards a specific target, it may be the case that the sentence will have opposing sentiments for different targets, this is another degree of complexity that targeted-sentiment analysis models need to account for as opposed to sentiment analysis of the sentence as a whole [**?**]. Phrases such as "great food but the service was dreadful!" convey different and opposite sentiments towards "food" and "service" [**?**]. Previous sentence oriented sentiment

analysis approaches such as [?], [?] would be incapable of correctly distinguishing this level of granularity [?].

[?], [?] also call attention to the fact that there are several instances where the sentiment conveyed by a particular word is contingent upon the target or aspect that is being considered. An adjective such as "short" can have positive connotations with respect to "waiting times" for a restaurant, on the other hand the same adjective is assumed negative when describing something such as the "battery life" of a product.

0.1.4 What metrics are commonly used to measure performance?

	$y = A$	$y \neq A$
$\hat{y} = A$	true positive (tp)	false positive (fp)
$\hat{y} \neq A$	false negative (fn)	true negative (tn)

Table 1: Confusion Matrix for the binary case of some class label, A. y represents the true label while $\hat{y}$ represents the predicted label.

Two commonly extrapolated metrics from which other measures are typically derived are precision and recall. Given some class A, the former is a ratio of correctly labelled instances to all instances labelled A whereas the latter compares the amount of correctly labelled instances to all instances of class A present in the data, which is analogous to the accuracy for class A. Formally, based on the definitions in table 1, the two measures are given by:

$$precision_A = \frac{tp_A}{tp_A + fp_A} \tag{1}$$

$$recall_A = \frac{tp_A}{tp_A + fn_A} \tag{2}$$

For the case with C classes, the total number of instances for a class c, N_c is equal to $tp_c + fn_c$. The prevalent metric for accuracy that is reported in the literature, equivalent to the micro-averaged recall, is thus computed by:

$$accuracy = \frac{\sum_c^C tp_c}{N} = recall^{micro} \tag{3}$$

Where N is the total number of instances in the data. Using this micro-averaged metric, however, is not necessarily the most accurate indicator of a model's performance in a classification task, particularly when the dataset that is being utilized is heavily biased to one specific class. Care must be given in the training phase of any machine learning model to ensure that the model is exposed to all classes in question in a balanced way. This is because in computing the micro-average, the weighting scheme is distributed across all instances in the dataset, as opposed to

the classes. A more sophisticated metric that is robust to this issue is the macro-averaged F1-score which equally distributes the weight across all classes as opposed to instances [**?**]. The macro-averaged F-measure is given by the harmonic mean of the, macro-averaged, precision and recall for a specific class. For some class A, this is given by:

$$F1_A^{macro} = \frac{2P_A^{macro}R_A^{macro}}{P_A^{macro} + R_A^{macro}} \tag{4}$$

Training a model heavily on one specific class, or not enough on another could lead the model to classify the majority of test samples to the biased class or being unable to correctly classify the class that has been under-represented in training, since the model would not have gathered enough information to discern this class. In the case where the testing dataset would be imbalanced towards the same class, the overall accuracy would lack the sufficient information expected as a metric to illustrate the effectiveness of the model to classify samples into the correct class since the model would have been trained in a biased way towards the class that is prevalent.

As an example, a frequently cited benchmark dataset is presented in [**?**], this dataset consists of 6248 training and 692 test phrases collected from twitter, each annotated with a particular sentiment (negative, neutral or positive) towards a specific target that appears in the tweet. Within both the training and testing subsets of this dataset, there are twice as many neutral instances as there are positive and negative instances. Works such as [**?**], [**?**] correctly point out the shortcoming of accuracy as a valid performance metric in this situations such as this, and cite macro-averaged F1 scores in their results.

0.2 Manual Feature Engineering

0.2.1 What did initial approaches using manual features involve?

Initially, the conventional approach involved manually extracting the most expressive and information rich features from sentences that would subsequently be processed through some statistical model such as a Support Vector Machine (SVM) for classification.

This entailed the formulation of processes by which to obtain these features, and was normally preceded by some form of normalization of the original data before these features could be extracted. Typically many types of these features were used in conjunction, each intended to extrapolate differing particularities about a specific aspect of the text, such as whether a specific token represented a noun or an adjective, or details about the words surrounding it to name a few.

0.2.2 What were some of the initial approaches using manual features?

The capacity of the SVM had been demonstrated on the general task of sentiment analysis in works such as [?], as well as other tools such as, bag-of-words, part-of-speech tags and other morphological and syntactical features and external resources such as linguistic parsers and sentiment lexicon, employed in works such as [?], [?], [?].

However, as [?] point out, these methods would implicitly impose an external dependency on the system. Moreover, within the context of social media, where conventional rules of language are often times regarded rather as guidelines, various studies question the applicability of dependency parsing techniques that rely on a particular degree of formality, or structure, within the content itself [?], [?]. Nevertheless these features have proven their worth when used in conjunction with powerful models such as the aforementioned SVM [?] [?], as well as neural networks [?], [?], in predicting sentiment polarity.

Even in the work that followed, focusing on increasingly autonomous feature extraction methods and more sophisticated deep learning architectures such as the Long Short Term Memory (LSTM) model, [?] make note of the competitive results obtained by the SVM approach in [?] when compared to their implementations.

0.2.3 What are the disadvantages of using manual features?

Although works such as [?], [?]) obtained encouraging results, much of the subsequent literature recognizes that these results where exceedingly contingent on the choice of features that were being utilized [?].

Although the manual feature-based approach fared well in their work, [?] suggest that features of this kind lack the required resolution of detail that would accurately capture the interplay between target and context. The features that had been used had sound rationales behind them, however devising these rationales was in itself becoming increasingly time consuming. One reason for this is scalability; with the increase of data that were available, this inevitably brings with it more considerations and specifics that must be accounted for when otherwise manually devising these feature.

As alluded to by [?], with the aforementioned increase in labor involved, these approaches were exhibiting diminishing returns, and could be regarded as a bottleneck in terms of performance of these models and the wealth of data available. A more autonomous solution that would accurately capture the intricacies of language from an expansive wealth of text at a deeper level, not contingent on a proportionally large amount of labor-intensive manual feature-engineering, was desired to further advance the field of targeted sentiment analysis.

0.3 Word Embeddings

0.3.1 What are word embeddings?

Word embeddings seek to tackle the non-trivial task of accurately capturing as much of the intricate details that are inherent in language, as possible. To model the intricacies of a word within the context of a language, or a particular subset thereof, sophisticated models are employed to construct continuous and real-valued vectors for each word. The resulting vectors are commonly referred to as word embeddings, and are meant to be numerical representations of contexts in which each word is used [?] [?] [?] [?].

By first learning a continuous word vector embedding from data [?] [?] [?], most approaches can take advantage of the principle of *compositionality* [?] to obtain a sentence or indeed a document level representation for a myriad of downstream NLP tasks, including sentiment analysis.

Two of the most prominent word embedding models that are currently employed in many NLP tasks are *word2vec* [?] and *GloVe* [?]. More recently, an extension on the former, termed *fasttext* [?] is also garnering considerable attention within the field, distinguishing itself from the other models mentioned through the use of sub-word information.

0.3.2 What are the leading two approaches?

The two principal methods for learning distributional word representations are count-based and prediction-based, each with their own strengths and shortcomings, and neither being clearly superior to the other in every aspect. The former typically involves performing dimentionality reduction on a co-occurrence count matrix, whereas the latter is derives word representations from learning to correctly predict words or contexts within a bounded, moving, window. [?]

Both methods provide a powerful means of autonomously extracting expressive and meaningful features from a wealth of text to the degree that would be unfeasible through manual means alone due to the staggering complexity inherent in language itself as well as the sheer volume of data that is being constantly made available through various online platforms that allow the power of expression to an unparalleled, and ever-growing, number of people across a myriad of different domains and topics.

One of the benefits that is had from constructing a vector space with distributional representations of words is the ability to model features such as similarity between the constituent words and subsequently group words with similar meanings which expands a downstream model's comprehension of a language. [?] Matrix factorization methods, such as Latent Semantic Analysis (LSA) [?] as cited in [?], attempt to extrapolate significant statistical information pertaining to a specific corpus, from decomposed matrices that are typically obtained through low-rank approximations. Although techniques such as LSA do this effectively, [?] note a

lackluster performance in word-analogy tasks, which ultimately depend, in large part, on accurately capturing the aforementioned similarity between words in the vector space.

Models trained using matrix factorization methods are efficient at exploiting statistical information, and are typically easier to parallelize albeit at a higher initial memory investment involved in storing the co-occurrence matrix [**?**]. [**?**] work on the assumption that more related words will tend to appear closer to each other than not, making window-based approaches ideal for word analogy tasks that benefit from accurate word-similarity modelling.

Since predictive methods rely on a bounded context window of some specified width when making predictions a trivial disadvantage that presents itself immediately with this approach is the limited capacity to efficiently exploit redundancy in the data on at a macro-scale. This is due to the fact that these methods operate on the level of their current context window as opposed to the document as a whole [**?**].

0.3.3 How does word2vec do its thing?

In their work, [**?**] note that the tendency in the field of NLP was to regard words in an isolated fashion as opposed to considering each word within the scope of a distributed space where more in-depth information could be encoded regarding the relationship between words. Models, such as the popular n-gram model where prevalent in the literature due to their simplicity and effectiveness. [**?**] note that these methods had largely peaked as unlike distributed representations, they were limited in their capacity to model relationships, such as similarity, between words.

There was a need for more sophisticated, and scalable, techniques to address the bottleneck that was presenting itself in the lack of accurately labelled data that was being made available for training models in fields such as Automatic Speech Recognition (ASR) and machine translation. As these models emerged to tackle increasingly large data sets, simpler models, such as the aforementioned n-gram model, were consistently outperformed by neural networks using distributed representations of words in the form of word embeddings. Moreover, neural networks were shown to capture the linear relationships between words more accurately than previous methods that used LSA, while maintaining a higher level of scalability when compared to LDA [**?**].

Some of the initial work making use of neural networks to learn word embeddings include [**?**], later refined in [**?**]. Their approach consisted of a feedforward neural network tasked with predicting the correct word within a context window of five words including the target word itself.

[**?**] introduced the Continuous Skip-gram and the Continuous Bag-Of-Words (CBOW) models in an effort to extract word embeddings from large corpora. As illustrated in figure 1, the former aims to learn an adequate representation of a word by training to predict the words that are most likely to surround it while,

inversely, the latter was trained to predict a specific word given its context, both using a feed-forward neural network with the non-linear hidden layer removed.

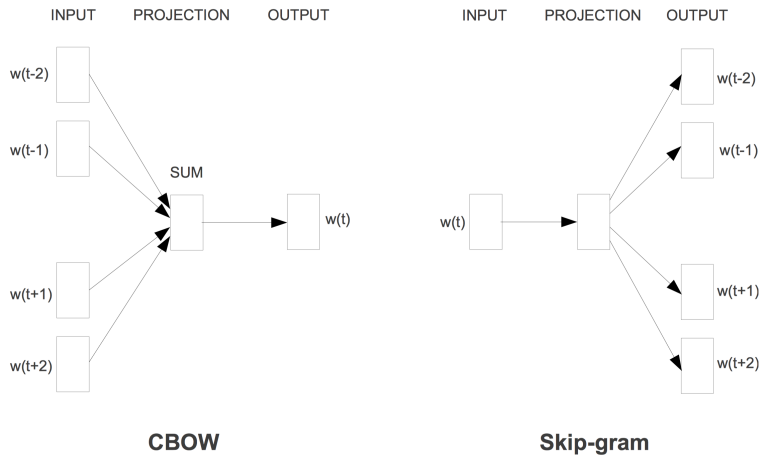

Figure 1: The differences between the model architectures proposed by Mikolov in [**?**]. The Continuous Bag-Of-Words (CBOW) approach predicts a word from its context and, conversely, the skip-gram model predicts the context from a word. [**?**]

Their work brought about a novel evaluation strategy which measures the expressive capacity of a word vector space through complex multi-dimensional comparisons that went over and above previous scalar metrics that were typically limited to the distance or angle between word vectors. As evidence of the level of sophistication obtained in the representations that were generated using the *skip-gram* model, the authors note that the composition of vectors for words such as "Germany" and "capital" results in a vector that closely resembles the word "Berlin". Additionally, more sophisticated linear translations could be modeled such as "Madrid" - "Spain" + "France" is resulting in a word vector closest to that of the word "Paris" [**?**].

While the skip-gram model required no dense matrix multiplications which made it substantially more efficient than most of the neural network implementations that had preceded it, in their experiments they note that the quality of the resultant word vectors could be improved by increasing the window size, however this would carry with it a corresponding increase in computational complexity. Nevertheless, [**?**] demonstrated that sufficiently expressive vector representations of words could be obtained from large corpora of data without the need for computationally expensive models.

Following their initial work, [**?**] later expanded on their *word2vec* models, introducing the negative sampling algorithm to improve the efficiency by which the model learned word vectors while proposing a method for accounting for phrases through a simple data-driven approach whereby particular phrases composed of multiple words were treated as a singular token, and trained-for as such.

As noted by [?], the performance of their models were contingent on a set of design choices, most critical of which include the training algorithm, vector size, sub-sampling rate and the size of the training window. They conclude that the optimal configuration for these parameters varies based on the task being tackled.

0.3.4 How is GloVe different to Word2Vec?

Using co-occurrence statistics to extract continuous representations of words within a large corpus of data has been explored in NLP in works as early as [?], as cited by [?]

[?] argue that while the significance of word occurrence information within a text when learning word representations in an unsupervised manner is uncontested, further research is needed into the mechanisms by which these statistical data generate meaningful vector representations. In pursuit of this, they propose the *GloVe* model, characterized by its use of the global, that is to say at the level of the corpus in its entirety, co-occurrence information to produce aptly-called *"global vectors"*.

[?] point out that while count and prediction based methods are not fundamentally dissimilar, since both exploit co-occurrence statistics within a corpus to obtain accurate representations, they argue for the efficiency of the former approach over the latter.

A word-word co-occurrence matrix X is constructed from a vocabulary such that X_{ij} is representative of the number of times word j is found in the context of word i. This invariably makes X sparse in nature, since the substantial portion of words within a language cannot be expected to occur within an equally substantial number of words.

[?] develop *GloVe* by only considering non-zero elements of the co-occurrence matrix of the corpus as a whole and as opposed to the sparse matrix in its entirety. This provides a substantial increase in speed and moreover, as their work suggests, generates more expressive representations of each word when compared to a limited window-based approach. *GloVe* was evaluated on three separate tasks, specifically word analogy, word similarity and entity recognition tasks, achieving superior results over the previous literature in all. [?]

Furthering the case for the capacity of the *GloVe* model, the comparative study carried out on the task of reading comprehension by [?] demonstrated that pre-trained *GloVe* embeddings outclassed other embeddings, including *word2vec* [?]. From their experiments, [?] continue to suggest that these embeddings, in their off-the-shelf format, also surpassed embeddings trained on the target text itself as well as, to their surprise, an expanded corpus of data extracted from a domain commensurate with that of the target text.

0.3.5 How is *fasttext* different?

Both *word2vec* and *GloVe* are considered as models operating on a word-level by regarding the word as the atomic operand, however [**?**] argue that this approach is possibly sub-optimal when considering languages, such as Turkish and Finnish, where single words can have multiple morphologies, or comprise of exceedingly large vocabularies, or both. In contrast, they argue that the use of sub-word information lends itself well to these languages where the multiple morphologies of a word follow some form structure, such as specific verb conjugations.

To tackle this issue they extend the *word2vec* skip-gram model to operate at a sub-word level, adopting a bag-of-characters n-grams approach for word representation. As opposed to having each word represented by a vector, a word is represented as the aggregate sum of its constituent character n-grams.

The authors also demonstrate how a vector for an OOV word can be constructed from its character n-grams with remarkable similarity to a comparable IV word. Some of their results can be seen in 2, where an OOV word "microcircuit" shows positive cosine similarity (in red) to an IV word "chip" between its constituent character n-grams "micro" and "circuit".

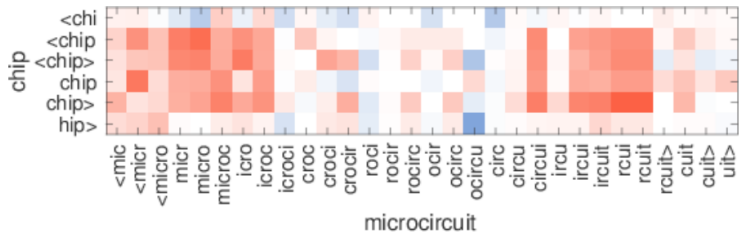

Figure 2: Character n-gram similarity between the OOV word "microcircuit" and IV word "chip". Positive and negative cosine similarity are denoted in red and blue respectively. Figure adapted from [**?**]

Need to include disadvantages of fasttext to make a contrast

0.4 Deep Learning

0.4.1 What are the fundamentals of a Neural Network?

Modelled on the human brain, neural networks typically involve a series of layers composed of neurons. Figure 3 illustrates one such simplified neural network architecture with a single hidden layer, L_2, preceded by the input layer L_1, and followed by the output layer L_3. (x_1, x_2, x_3) represents a three dimensional input vector, whereas the neurons, or hidden units, of the of the network are depicted as (h_1, h_2, h_3). The firing action of each neuron is expressed using a non-linear activation function. This action propagates from one neuron in a layer to all other

connected neurons in the subsequent layer and is modulated by a particular weight that characterizes each intra-neural connection. At minimum, these networks will incorporate an input and an output layer that will encapsulate one, or more, hidden layers.

The network is able to learn, or model, a function by adjusting the weights to minimize a some measure of error towards a particular objective function. [?]

The hidden layers of a neural network architecture are able to extrapolate features at a level that cannot be carried out manually, thereby capturing more subtle details and generating richer representations of the data being learned. Foregoing this need for extensive manual feature engineering is one of the primary drivers of neural network models, even though these are approaches are typically black-box solutions, with obfuscated inner-workings.

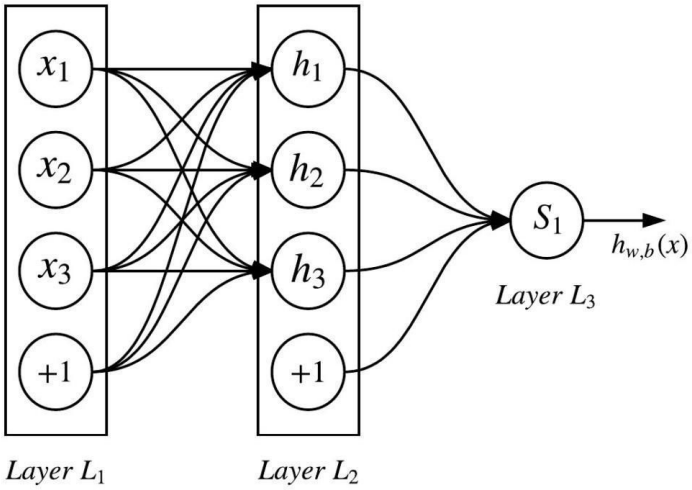

Figure 3: Standard feedforward neural network (FFNN) architecture [?]

0.4.2 What are the common activation functions?

Three of the most commonly used activation functions are sigmoid (eq. 5), hyperbolic tangent (eq. 6) and Rectified Linear Unit (ReLU; eq. 7). Although the choice of activation function is typically heuristic, [?] as cited in [?] notes that the ReLU is easier to compute when compared to the other two and also tends to converge faster, all while maintaining similar or even better performance.

$$f(W^t x) = sigmoid(W^t x) = \frac{1}{1 + e^{-W^t x}} \tag{5}$$

$$f(W^t x) = \tanh(W^t x) = \frac{e^{W^t x} - e^{-W^t x}}{e^{W^t x} + e^{-W^t x}} \tag{6}$$

$$f(W^t x) = ReLU(W^t x) = \max(0, W^t x) \tag{7}$$

0.4.3 How are Neural Networks typically trained?

Neural networks are trained by following the direction of the gradient that lessens the measured error rate of some objective function which is calculated through the repeated application of the chain-rule. The process can be carried out on the training set in its entirety, commonly referred to as batch learning, or in an on-line manner, carrying out updates after each training sample. The latter is known as stochastic gradient descent and is known to be more efficient and robust to local minima when dealing with large datasets [**?**] as cited in [**?**].

There will always be noise in the sampled training data that does not translate to the real world test data scenarios and that the model must therefore avoid learning. Training models to the point of becoming overly-sensitive to this noise is referred to as over-fitting the data, and there are a number of regularization techniques often employed in the literature while training to counteract this.

One common approach is introducing some form of weight penalties, for instance, $L1$ and $L2$ regularization. Other techniques frequently used include *early-stopping*, whereby a fraction of the training data is used as a validation set that the model is tested against at regular intervals during training to test for a sustained improvement [**?**], and *dropout* [**?**], in which random neurons in a NN are ignored ("dropped") with some probability p, effectively training a diverse range of "thinned" versions of the original NN. An approximated average of those networks is subsequently used as final trained model by scaling its weights by that same dropout probability p.

0.4.4 How are Deep Neural Networks formed?

"Deep" neural networks are constructed by stacking a series of layers in such a way that salient features extracted from one layer are passed on as input data to the next layer, and so on. This stacking process, in theory, improves the capacity of the network to extract more abstract and expressive features that reside at a deeper level within the input data. [**?**]

0.4.5 Why is Deep Learning popular again recently?

In recent years, the astounding progress in computing resources such as GPUs and high performance distributed computing have made deep learning accessible on an unparalleled level. Consequently, this has driven interest in the applicability of deep learning architectures such as the CNN and RNN in a range of fields from computer vision to natural language processing [**?**], [**?**].

0.4.6 What are two common deep learning models used? (CNN and RNN)

Not least owing to their impeccable ability for effectively extracting highly expressive low-dimensional representations of text automatically, coupled with the aforementioned resurgence of deep learning due to the increase in available computing power, the use of neural network models such as [?] [?] [?] became increasingly widespread within the field of NLP, including targeted and aspect based sentiment classification tasks [?]; [?]; [?] [?].

The most prevalent architectures are the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNN) and the descendents thereof. Other deep learning models worth mentioning include the Recursive Neural Network (Rec-NN), which has been employed in works such as [?] and [?] for syntactic analysis and sentence sentiment analysis respectively. [?]

0.4.7 What are the fundamentals of a CNN (in brief)?

The layers typically comprise of filters, also referred to as kernels, of differing widths that slide over the input data to extrapolate various features. Each of these features would be representative of a diverse set of aspects of the data that would aid in defining it with respect to the task being tackled.

These convolution layers are coupled with a max-pooling layer with the intention of extracting the most salient values. These values can subsequently be forwarded to another convolution layer that presumably extrapolates deeper, more abstract features. This process can be repeated a number of times across multiple convolution layers to build a deep CNN that eventually produces a single feature vector representation of the original input data. Figure 4 illustrates this process using six filters with three difference widths and two diverse operations for each, the results of which are down-scaled using max-pooling, and subsequently passed through a softmax function for binary classification.

0.4.8 What are the advantages of using max-pooling?

Max-pooling is ideal for producing a fixed-length representation of the data, which is a common prerequisite for classification tasks, while at the same time preserving the most prominent features from the original data. [?]

0.4.9 What are some approaches using CNN?

Some of the first CNN-based approaches that paved the way for the growth of CNN architectures in the literature that followed were [?], [?], and [?]. [?] used CNN architecture for sentence classification tasks ranging from subjectivity and question type with promising results albeit in the face of a number of challenges that became evident to the authors. Not least of these was the limited capacity for the CNN

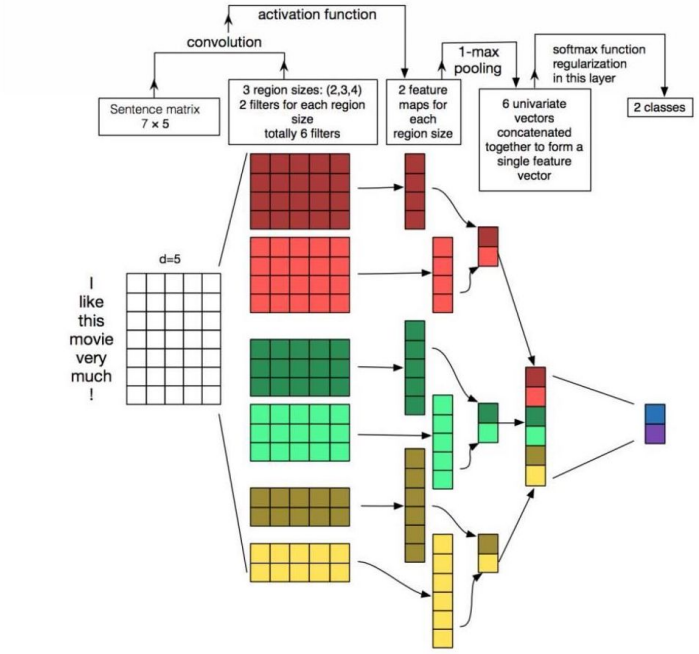

Figure 4: An example of a CNN architecture used for sentence modelling and subsequent binary classification [?] as cited in [?].

architecture to capture syntactical dependencies in sentences that occurred over long distances. This challenge was the one of the foremost drivers for the work that followed by [?] who developed the DynamicCNN (DCNN) which consisted of a series of hierarchical convolution and k-max pooling layers.

Other works cited in [?] include [?] that dealt with sarcasm detection on twitter data, noting a need for additional context information when dealing with short texts of this nature. This observation is also echoed in [?], who noted superior performance of CNN networks when dealing with longer text that would provide more contextual information as opposed to shorter text. Due to the vast amount of parameters that CNNs typically need to learn, scarcity of data is an often cited challenge [?].

Work carried out by [?] made use of a CNN to deduce the sentiment of the target based on the sentiment of the clause surrounding it. As noted by [?] however, this method still operated on the assumption that the most salient features of a word can be extracted from other words in close proximity.

A phrase such as "I bought a mobile phone, its camera is wonderful but the battery life is short, not particularly satisfied overall", challenges this assumption with respect to the "mobile phone", as the intended target since the most sentimentally-laden words appear at the opposite end of the sentence.

0.4.10 What are the fundamentals of an RNN based model (in brief)?

A Recurrent Neural Network (RNN) can be thought of as a chain of recurring modules which typically represent elements within a variable-length sequence, with an internal hidden state that represents the network's "memory". This state is passed forward from one module in the chain to the next.

Unlike a CNN, where each layer has its own set of trainable parameters that must be learned, a RNN uses a single set of parameters across all of the modules in the chain which significantly diminishes the total number of parameters that it must learn.

Through forwarding the hidden state from one time step in the chain to the next, the network is able to "remember" information from previous elements of the sequence and use that information when generating a representation for the current element [**?**].

0.4.11 What makes the RNN more suitable for targeted sentiment analysis? (sequence)

The hidden state that characterizes RNNs acts as its "memory" element and makes these networks particularly effective in dealing with data that are sequential in nature. One of the most prominent examples of these data is language, where the significance of a word at one time step may be substantially altered by those that preceded it. Consider, for example, the word "dog", for which the meaning would shift entirely, from an animal to a popular American snack, should it be preceded by the word "hot" [**?**].

Moreover, the capacity of RNNs to model variable length sequences to a fixed length representation also makes them particularly practical when dealing with different units of resolution in languages including documents, sentences, and even words, all of which are naturally arbitrary in length. [**?**]

0.4.12 What is problem faced by RNN? (vanishing & exploding gradient)

In the setting of a traditional RNN, the tendency for a gradient to exhibit exponential change to the degree that prevents substantive learning increases with the length of a sequence [**?**] [**?**]. This phenomenon is what is implied by the terms "vanishing" or "exploding" gradients, where the changes observed over time steps often vanish with time or, although less frequently, but with equally devastating results, grow exponentially; hindering the network's capacity for learning.

0.4.13 What solutions exist to vanishing/exploding gradient problem? (LSTM and GRU)

One class of solutions to the vanishing or exploding gradient problem is to carry out particular modifications on top of the standard stochastic gradient descent algorithm, such as gradient clipping, whereby the norm of the gradient vector is *clipped*, or using second order derivatives, which may or may not be influences to a lesser degree. [**?**] The second, and more popular, class of solutions look instead to introduce further sophisticated, additive, non-linearities to the traditional RNN unit. These would selectively carry forward salient features and filter out irrelevant information from previous time steps, as opposed to overwriting the memory content at each time step.

Variants on the standard RNN network were proposed to address the vanishing and exploding gradient issue. The most popular of these being the Long Short Term Memory (LSTM) and, more recently, Gated Recurrent Unit (GRU). Other approaches include, but are not limited to, Residual Networks (Res-Net).

0.4.14 What are the fundamentals of the LSTM (in brief)?

LSTM [**?**] is an extension on the RNN model that addresses the issue of a vanishing or exploding gradients through the use of gates that control the flow of information from the past states to present states.

To do this, the LSTM introduces three adaptive gates that can be considered additional neural layers on top of the single neural layer that characterizes the typical RNN. The layers introduce an increased level of sophistication in the "remembering" process from one sequence point to the next.

These gates are commonly referred to as the input, forget and output gate. Moreover the LSTM also maintains a two inner states as opposed to one, namely the cell state and the hidden state.

The process undergone in each repeating module of the standard LSTM architecture depicted in figure 5 starts at the forget gate. Here, the information to be removed from the cell state C_{t-1} is selected through the sigmoid layer (eq. 8).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{8}$$

Next, an update vector is produced through a point-wise multiplication operation between the input gate (9), which represents the values selected to be updated, and a vector of candidate values, $\tilde{C}$ (10). This update vector is added to the current cell state, from which the forget gate had previously removed information deemed redundant.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

$$\tilde{C} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{10}$$

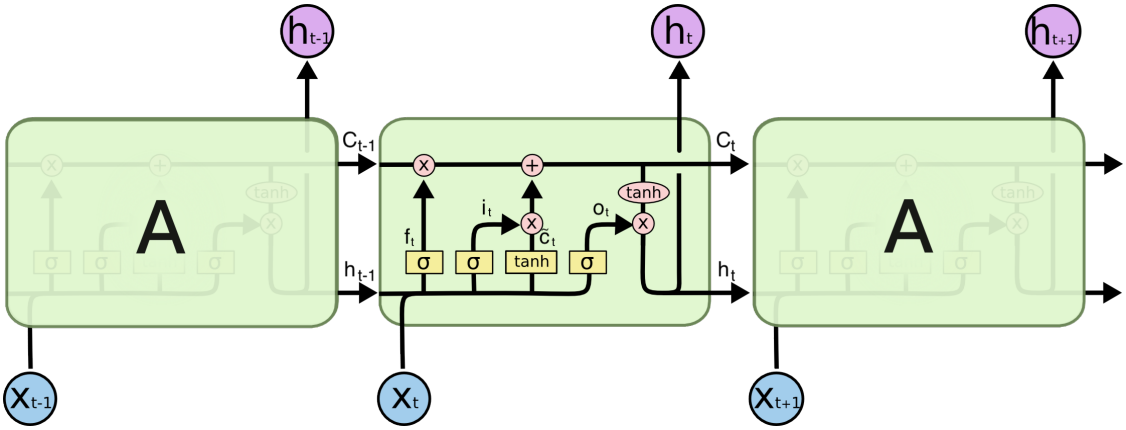

Figure 5: LSTM repeating module illustrating the four neural layers (Yellow Boxes) that comprise it. Point-wise vector operations are depicted in red boxes. Image adapted from [?]

Finally, an output gate (11) regulates the amount of cell state information that is output to the rest of the network through the unit's hidden state h_t (eq. 12).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t * \tanh(C_t) \tag{12}$$

0.4.15 How does the LSTM model address the vanishing & exploding gradient problem?

The gating mechanism that is present in the LSTM network allows for the persistence of salient features that are encountered early in the sequence and which would otherwise be overwritten in a typical RNN architecture.

The input and output gates of the LSTM control the amount of memory content that is to be added to the memory cell and the memory content that is exposed to the rest of the network respectively.

Forget gates were later added to the architecture [?], which enabled the memory cells to reset themselves. It is important to note that the LSTM updates the memory cell independently from the forget gate, which is to say that, for the LSTM to "remember" a prior input in a sequence two conditions must be satisfied; the input gate must be closed, preventing the addition of new information and, the forget gate must be maintained open, as otherwise this would cause the existing memory content to be reset.

0.4.16 What are some approaches using LSTM?

[?] achieved results competitive with those obtained by [?] by using an LSTM, as opposed to CNN, to generate representations of tweets. [?] notes that their model

was able to extract features over long distances at a fraction of the complexity of the DynamicCNN [**?**].

Another variant on the LSTM architecture involved the use of dependency parsing to build tree-structured LSTMs which operate on some specific tree-pattern that is extracted from the data, typically through the use of external parsing tools [**?**]. Approaches of this variety [**?**], [**?**], [**?**] have obtained promising results, however as [**?**] note, these are contingent on the data being well-formed structurally and grammatically, which is not guaranteed when dealing with micro-texts on social media platforms.

Two of the foremost extensions on the classical LSTM model when tackling target-based sentiment analysis, were proposed in [**?**] with the intent of accounting for the target information, these were termed the Target Dependent LSTM (TD-LSTM) and Target Connection LSTM (TC-LSTM). Their work showed that the changes proposed would improve the performance over a standard LSTM. [**?**] generate representations of a target's left and right contexts, where the target representation is concatenated to each. These two are then finally chained together to form the final, target-specific, representation of the sentence which is fed to an LSTM for sentiment classification [**?**].

TC-LSTM [**?**] was reported as one of the first neural network based methods to obtain state-of-the-art performance without the use of laborious feature engineering and external data sources. It is worth noting however that recent work [**?**] failed to reproduce the original results that were reported.

0.4.17 What are difficulties mentioned that LSTMs deal with?

While the issue of long-distance dependencies and CNN based approaches is often brought up, [**?**] remarks that LSTMs will still struggle with features that are located too far apart within a sequence, given the phrase "Except Patrick, all other actors don't play well", an LSTM would struggle to identify the positive sentiment on the opinion target "Patrick" due to the distance between the terms "Except" and "don't play well".

From the observation made by [**?**] when employing LSTM-based models in the field of machine translation, [**?**] make the case that the TD-LSTM would suffer from in instances where the most sentimentally salient word is located further away from the target being considered.

0.4.18 How does a Bi-LSTM differ from an LSTM?

When dealing with bounded sequences, a natural conclusion one might draw is that valuable information can be extrapolated from future contexts as well as past contexts. Bidirectional RNNs are employed with this specific intention in mind.

Building on top of the LSTM architecture, a Bi-LSTM is so called as it is

the result of two stacked LSTMs, each responsible for processing and extracting features from a sequence in opposite directions. Features extracted from these two directions are then typically concatenated into a single, theoretically richer, and more expressive, representation of the sequential data.

It has been argued that the bidirectional nature of these models violates causality, which is justified to an extent, since a future context cannot be exploited if it is the same element that is being predicted. An example of this is predicting future stock market fluctuations. However, for tasks where this information is available, bidirectional variants that make use of this information have consistently been shown to outperform their uni-directional counterparts. [**?**]

0.4.19 What are some approaches using Bi-LSTM?

Sentence and document level sentiment classification tasks lend themselves well to the use of bi-directional RNN (BRNN) variants, since the boundaries of the sequence being classified are well-defined a priori. In their work, [**?**] suggest that using a BLSTM improves the ability of their approach to extrapolate phrase-like features when compared to a uni-directional sequential approach.

An improvement in results stemming, in part, from taking a bi-directional approach is also observed in [**?**], in both accuracy and macro F1 scores when compared to the work of [**?**], from which the former was inspired. More recent works (eg. [**?**], [**?**]) that have reported notable results in the field also take advantage of bi-directionality in their approach.

0.4.20 What are the fundamentals of the GRU (in brief)?

The GRU does away with the cell state that is found in the LSTM, maintaining only a single hidden state as its "memory". The second way in which the GRU differs from the LSTM is in its number of gating units, by foregoing the output gate and combining the input and forget gates into a single update gate, the GRU benefits from having less parameters to learn.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{13}$$

As the name implies, when the reset gate (eq. 14) nears 0, it allows the GRU unit to drop the previous information, which may be deemed inconsequential at a later time, and reset itself with the current input only. This information is then used by the GRU to produce a vector of candidate activation values (eq. 15).

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{14}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t] + b) \tag{15}$$

Instead of maintaining an internal memory cell, the GRU uses the update gate and carries out a linear interpolation function (eq. 16) to control the amount of

candidate activation, $\tilde{h}_t$, that is added to the previous activation. In this way the update gate serves as the primary mechanism preventing the GRU from overwriting a previously encountered salient feature.

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{16}$$

Whereas the LSTM utilizes the output gate to control the exposure of its internal memory to the rest of the network, the GRU does not have any means by which to control the exposure of its inner state, and therefore exposes its hidden state in its entirety to the rest of the network.

Each unit in a GRU model will have separate update and reset gates that will independently learn to capture long and short term dependencies in a sequence. Units that learn to capture longer-term dependencies will have active update gates whereas, for short-term dependencies, the reset gates will tend to be more active.

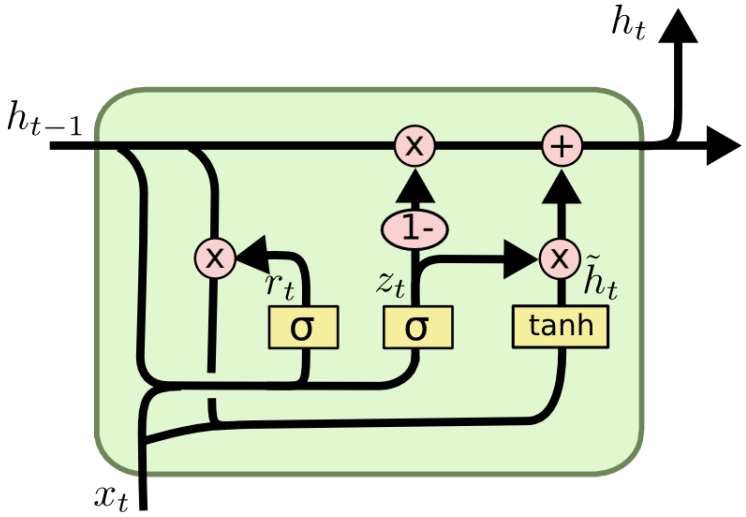

Figure 6: Internal gating structure of a GRU unit. [?]

0.4.21 What are some approaches using GRU?

[?] model the interplay between targets and their contexts through gated mechanism operating on the a representation of the original sentence split into three components using a gated-RNN.

[?] use a GRU as part of their approach, serving as a gating mechanism to determine whether to update a specific part of memory based on past activations. They claim state of the art results in aspect detection and sentiment classification, outperforming SenticNet [?], and foregoing the need of external knowledge.

[?] use gated tanh-ReLU units to control flow of features from convolutional neural networks to max pooling layer. They deal more with the advantages that a CNN offers, since it is not time-dependent and is therefore easier to parallelize.

[?] generate a sentence representation by stacking two GRU based networks, similar to the BLSTM configuration, opting to go with the GRU architecture instead since it has less parameters to learn, noting similar results. A continuous vector representation of the target is sandwiched between the two GRU direction outputs, and fed to softmax classifier. Authors report accuracy and macro-f1 scores better than state of the art obtained on a twitter dataset [?].

0.4.22 What advantages/disadvantages does the GRU have over the LSTM if any?

Since the architecture of the GRU is less complex than that of the LSTM, there may be circumstances where the former may be more efficient than the latter since it requires less parameters to learn [?], [?].

That being said, there is no clear winner between the two variants, as demonstrated by work conducted in [?]. The results obtained, shown in figure 7, concluded that the only demonstrable advantage is that of both variants over a standard RNN architecture. The authors go on to say that the performance of the two variants themselves is contingent on the particular nature of the tasks being addressed. Although it is worth noting that the work carried out by [?] was not specifically in the field of NLP, the deciding factor between these variants in NLP literature still tends to be heuristic [?].

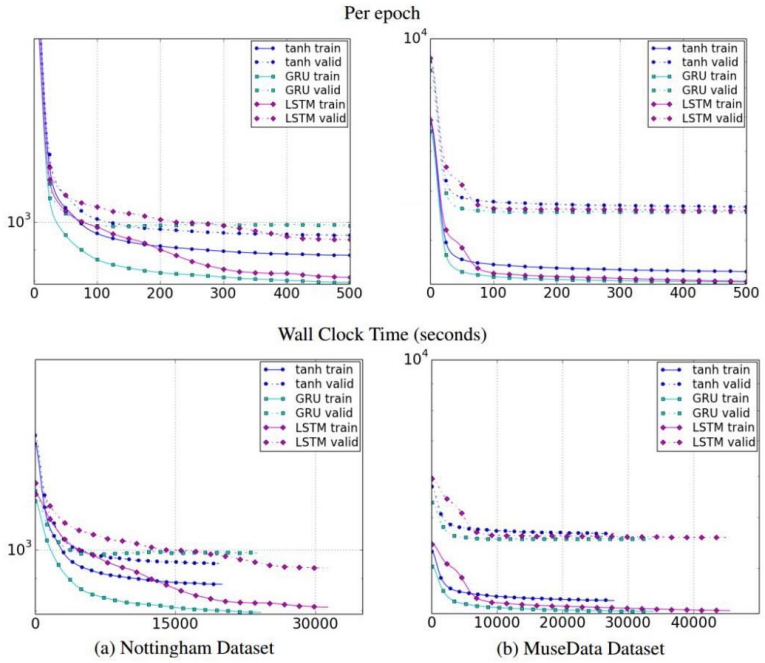

Figure 7: Results obtained by [?] during training and validation of different RNN variants illustrating the superiority of LSTM and GRU units over the traditional RNN.

0.4.23 Are there any advantages of CNN over RNN models?

In spite of various studies illustrating the fact that CNN networks are data heavy in nature and often require auxiliary data when dealing with micro-texts such as those obtained from social media networks such as twitter, [?] note that assuming the superiority of RNNs for sequential data is inaccurate.

They proceed to cite studies where CNNs performed competitively, and even surpassed RNN-based architectures in tasks for which the latter would, in theory, be better suited, such as language-modelling [?] as cited in [?].

It is worth keeping in mind that when dealing with language modelling, RNNs and CNNs approach the task from separate avenues; while RNNs, and their descendents, have no context boundaries when processing sequential data to generate a representation, CNN-based approaches seek to extrapolate the most meaningful n-grams from bounded-contexts from that sequence, to produce their final representation. [?]

0.4.24 Are there any approaches that use both CNN and RNN models?

There have been works in the field of sentiment analysis, targeted and otherwise, with the intention of taking advantage of the benefits of both CNN and RNN based models in an ensemble approach. *Finki* [?] and *BB_twtr* [?] are two solutions of this nature, dealing with sentiment analysis of tweets. These models involve the coupling of LSTM or GRU models with CNN models to carry out both binary, and fine-grained (five degrees) sentiment analysis on tweets. The works were part of the submission made to the yearly SemEval conference, and each performed impressively in their respective rounds.

While [?] and [?] were not addressing *targeted* sentiment analysis, recent works such as [?] and [?] also marry GRU or BLSTM with CNN models. Interestingly, both works cite the simplicity of their approaches when compared to more complex attention mechanisms as a driving factor, while still achieving comparable results in their experiments.

0.4.25 What does an attention mechanism seek to model?

The intuition behind an attention mechanism is that, when dealing with sequential data, different parts of the sequence contribute to varying degrees towards a specific task or goal. This is abundantly apparent in the field of machine translation, where the attention mechanism made its debut [?]

When translating a particularly long sequence, for example, it is natural to focus principally on specific regions related to the current element being translated, as opposed to the source sentence as a whole, as this would result in the most relevant parts possibly being shadowed by unrelated material.

0.4.26 Where was attention first applied?

The remarkable power of attention mechanisms for machine translation [**?**] sparked an interest in investigating their applicability in a range of other tasks, including target, and aspect, based sentiment analysis [**?**] [**?**] [**?**] [**?**], image captioning [**?**] and question answering [**?**], where the notion of the most salient features being dispersed unevenly across the source data also holds true.

0.4.27 What are the fundamentals of the standard attention mechanism?

In its simplest form, an attention mechanism involves a layer which produces a weight vector that represents a distribution of salience across an original feature vector, which would otherwise be considered evenly in its entirety, with respect to some target. The nature of the target varies depending on the downstream task, in targeted or aspect based sentiment analysis this is typically the target or aspect in question, whereas in fields such as machine translation this could be the last hidden state that was output.

This process typically takes the form of some scoring function, such as a simple FFNN, which can be trained alongside the rest of the model, followed by a softmax layer. Given a series of representations $[h_1, h_2, ..., h_n]$ and some target t, the scoring function a calculates the salience of each h_i with respect to t. Subsequently, the softmax layer squashes the attentions scores into a valid distribution vector α with values in the range $(0, 1)$.

$$\alpha_i = \frac{\exp(a(h^i, t))}{\sum_{j=1}^{n} \exp(a(h^j, t))} \tag{17}$$

This weight vector is then used to produce a context vector c, as the weighted sum of the original feature vector. This process will amplify features with high attention weight values (approaching 1) while attenuating features with low attention weight values (approaching 0).

$$c = \sum_{i=1}^{n} \alpha_i h_i \tag{18}$$

0.4.28 What are some approaches that implement attention?

ATAE-LSTM [**?**] was one of the first to incorporate attention into an LSTM model for the purposes of aspect-based sentiment analysis. A vector representation of the target aspect is used as the subject of attention, allowing the model to attend to different parts of a sentence with respect to the aspect.

[?] make use of multiple attention layers to extrapolate the most salient, and sentiment-bearing words with respect to a target, suggesting that through more than a single attention layer, the model would be more effective at extrapolating features over long distances. They subsequently aggregate these attention results non-linearly using a GRU. The authors attribute their preference of a GRU over an LSTM for this stage in the process due to the former requiring less parameters than the latter.

[?] argue that prior works had focused primarily of the representation of contexts and not on targets themselves. When considering targets that consist of multiple words, the idea that words should not necessarily contribute equally to the final representation of that target is a valid assumption to make. They cite the example of "picture quality" as a target and argue that in such a case the word "picture" would play a more important role that "quality". To address this, their Interactive Attention Network (IAN) incorporates two attention networks to model both the target and the context interactively, to obtain a representation of the effect each had on the other.

More recent work, by [?], build on the intuition of producing better target representation as well as context representation. Their LCR-Rot model is characterized by a novel "rotary attention mechanism" that attempts to better capture the interplay between a target and its context as well as the contexts on the target. They argue that left and right contexts affect the target representation to a degree that merits a separate representation of the target for each, one which is "left-aware" and another that is "right-aware". [?] demonstrate the effectiveness of their rotary engine approach citing state-of-the-art results in accuracy on three distinct datasets, suggesting that properly modelling the effect of the target on the context may be as important as that of the context on the target.

0.4.29 What is the rationale behind memory networks?

It can be argued that one of the most prominent contributions of the attention mechanism covered in the previous section is the ability to selectively read information, typically from the internal state of a model, in a differentiable manner by reading from all of the data, to varying degrees. This, however, has more profound implications: the same process can be applied to selectively write in a differentiable way. Memory networks are so-called as they exploit this fact, providing an external memory store that a model can learn to refer to and update over time.

0.4.30 What were the first works that employed memory?

This concept was first actualized in two distinct, yet coincident, studies: [?], who proposed their Neural Turing Machine (NTM) and [?] who put forward their Memory Neural Network (MemNN) framework.

0.4.31 What are the fundamentals of memory networks?

In their seminal work, [?] describe their memory network framework as comprising of some tangible memory component, which is represented as an encoded continuous matrix. This external memory is updated through the use of neural network operations which selectively read and write to and from it. They proceed to conceptualize these operations in the form of four fundamental components.

The first component, the input component, I, is tasked with mapping incoming data to an internal representation. Pre-processing and embedding look-ups are two examples of operations that may be entailed at this stage. A generalization component, G, follows, which uses the data from I to update the memory. G is so-called as it allows for the potential of generalization of existing memory towards some future goal. In its simplest form however, this component simply stores incoming data in the next available memory slot, leaving existing memory unchanged. Using the input data from I and the current memory state an output component, O, produces an output feature vector which typically involves inferring the most relevant memories. This output is finally interpreted by a response component, R, to the desired format.

Each component that makes up this process may represent any trainable model such as an RNN or SVM, and trained accordingly. In their original approach, [?] use "hard" attention when probing for the most relevant memory evidences, where the number of highest scoring evidences is a tunable parameter. [?] build upon this idea by opting instead to use a softmax operation, which is conceptually comparable to using a "soft" attention mechanism over the external memory store. Moreover, unlike its "hard" counterpart, this makes the process differentiable, making the model trainable in an end-to-end fashion, requiring less supervision when compared to [?].

This framework put forward in [?] and subsequently extended by [?] served as the foundation from which most memory-based targeted sentiment analysis approaches emerged.

Furthermore, [?] show that the performance of their model can be enhanced by having the O component repeatedly attending to memory for a number of consecutively stacked "hops". Indeed, this observation is echoed in subsequent studies inspired by their work in the field of targeted sentiment analysis (eg. [?], shown in figure 8), the intuition being that through these consecutive hops the model is able to attend to richer, more abstractive, features than those existing solely on the surface level of the data.

0.4.32 What are some targeted sentiment analysis approaches that use memory networks?

Memory networks were first put forward towards the goal of targeted sentiment analysis by [?] (figure 8). Their approach was inspired by [?], with some differences in the attention function that was used, opting instead to use the method put

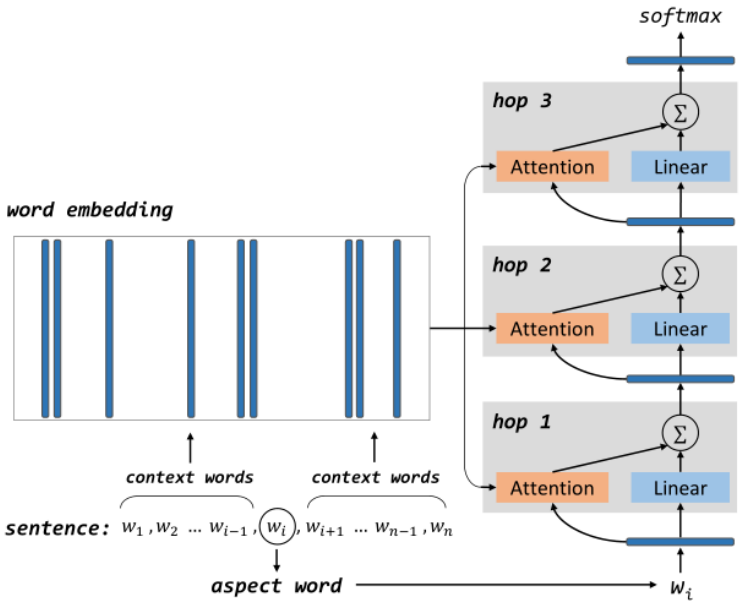

Figure 8: The deep memory network approach for targeted sentiment analysis, with 3 hops. The model stored the context of a target in memory and repeatedly attends to that memory with respect to the target word w_i. [?]

forward in [?]. The authors reported results outperforming the state-of-the-art SVM-based model by [?], which required extensive manual feature engineering, absent from their proposed memory network, as well as the far more complex LSTM based models [?] which required substantially more time to train. Similar to [?], the authors noted a marginal increase in performance as they increased the number of computational hops, which capped at around 8 hops.

[?] construct a location-weighted memory module from the hidden states of a BLSTM placed between the input and the attention modules so as to better capture information from phrases comprising of multiple words, such as "not wonderful enough". Moreover, unlike the [?], they combine the results of their recurrent attention layers non-linearly. These improvements led to a performance boost over [?]. The proposed model architecture is illustrated in figure 9. It is also worth pointing out that the authors failed to reproduce the results originally cited in [?].

[?] maintain a memory chain of entities that are encountered while processing a phrase and develop a gating mechanism that determines how those memory chains should be updated. The gating mechanism accounts for the content and the location of the memory chains compared to the current input, as well as the past activations, through the use of a GRU unit, when choosing to update a particular memory element.

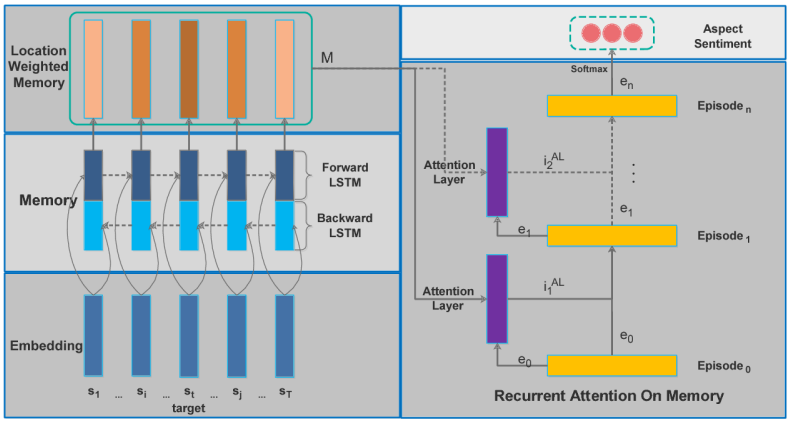

Figure 9: The recurrent attention model architecture showing the placement of the location-weighted memory module with respect to the input and attention layers. As opposed to [?], the attention values are combined from one episode to the next non-linearly, using a GRU. [?]

0.4.33 Why is it imperative to attend to targets as well as contexts?

Recently, work carried out by [?] showed an interesting performance boundary on approaches to targeted sentiment analysis that make use of attention mechanisms. In particular, [?] note that when context words have diametrically opposing sentimental bearings based on the target being considered, this cannot be modelled by improving by attending to the context alone.

To illustrate this, [?] consider two different targets, *price* and *resolution*, in four different phrases; "high price", "low resolution", "high resolution" and, "high price". Some sentiment score, s, for these phrases, where $s < 0$ implies a negative sentiment and $s > 0$ implies a positive sentiment accordingly (eq. 19). The attention weight α will have a value of 1 since the context consists of a single word which is represented by h. v represents the target and W is the weight matrix the model must learn.

$$s = W(\sum_{i=1}^{n} \alpha_i h_i + v) = W(h + v) \tag{19}$$

From equation 19, the following inequalities can be obtained,

$$\begin{aligned} W(h_{high} + v_{price}) &< 0 \\ W(h_{low} + v_{price}) &> 0 \\ W(h_{high} + v_{resolution}) &> 0 \\ W(h_{low} + v_{resolution}) &< 0 \end{aligned}$$

Expanding these inequalities results in $Wh_{high} < Wh_{low} < Wh_{high}$, which is a contradiction, and can therefore not be learned by the model. [?] point out that this

condition cannot be rectified through further attending to the context but rather ameliorating the representation of the targets v to better capture the complex relationship they have with the context in a way that possibly reverses the polarity of the final sentiment score s.

Towards this end, [**?**] experiment using a series of techniques, and note an improvement over previously cited results as well as a direct improvement on the RAM model [**?**] when these are coupled with their optimally performing target-sensitive context modelling strategy.

0.5 Out-of-vocabulary Words

0.5.1 What are OOV words?

A common theme in the deep learning models that have been discussed herein is the uncanny similarity between the techniques that these models employ to understand information and the same techniques that the human brain takes advantage of, both consciously and unconsciously, to produce an understanding of the information it is presented, with respect to a particular goal.

Traits such as properly identifying the most important and informative elements of a sequence, and referring to past events when making such decisions as well the fundamental process by which the intricacies of the networks are tuned when errors are made in training, in an effort to lessen such errors in the future.

In this analogy, when dealing with language-related tasks such as targeted sentiment analysis, the word embedding matrix from which different words' are represented numerically can be regarded as the working vocabulary that the model has prior to tackling the task at hand.

It stands to reason that due to the ever changing and ever evolving nature of language, it is impossible to account for the entire vocabulary of a particular language when constructing word embeddings. Regardless of the amount of text that is initially used to construct the word embeddings there shall be words that are not encountered within that text and therefore a continuous vector representation of that word could not be produced. While the probability of covering the most commonly occurring words increases with the size of the original text and the variety within it, encountering new words is an inescapable eventuality. These previously unseen words are consequently referred to as out-of-vocabulary (OOV) words.

The magnitude of the challenge that OOV words present, and the potential repercussions thereof, become evident in the landscape of this analogy. These represent words which the models essentially have no understanding of and can be of little help to it when attempting to extract any information it may convey to the task at hand. While the model can be expected to learn more about these words through repeated encounters in different contexts, the fact that these words, by their own nature, tend to occur infrequently in language substantially diminishes

the efficacy of this learning process.

0.5.2 What problems for generative tasks do OOV words create?

OOV words are of particular concern when dealing with tasks that are generative in nature, such as ASR. The toll of OOV words on the performance of approaches to these tasks is two-fold. Firstly, OOV word may be substituted with an incorrect IV word. Secondly, the OOV word has a direct effect on the neighbouring IV words [**?**].

A common approach to this problem is clustering OOV words into groups that would are sufficiently expressive of their constituents. Various techniques have been employed towards this goal such syntactic and morphological features, part-of-speech tag information, online resources, and subword-level models to name a few, [**?**] does a good job of outlining these approaches.

0.5.3 What problems for sentiment analysis do OOV words create?

Since sentiment analysis is classification task, where words are provided as input and subsequently used as keys when looking up the relevant embedding vector, substituting an OOV word for an IV word is of no concern. The effect of an OOV word on its neighboring words, however, is prone to undermine a model's ability to generate an accurate representation of the content as a whole.

This sort of phenomenon is not particularly difficult to imagine since it is often times the case in languages that a single word can have drastic effects on the meaning of a phrase, particularly in situations expressing negation. Consider a phrase such as "It avoids all the predictability found in Hollywood movies.", where "predictability" conveys a negative sentiment, which is subsequently negated by the verb "avoids".

Moreover, OOV words obviously make the process of comprehending a phrase more difficult by introducing elements that the model has no knowledge of. If the word embedding model that is being used is analogous to the model's understanding of a language, an OOV word is effectively a word the model does not understand, and therefore has limited means by which to gauge the effect of that word on the overall sentiment of the phrase, if any.

0.5.4 How are OOV words typically regarded in Sentiment Analysis?

A typical approach to this OOV challenge within the field of sentiment analysis is the use of a particular singular token that is meant to represent low frequency words during the training phase, and subsequently model all OOV words encountered in

the test phase. The vector for this token is often times initialized using some bounded random uniform distribution.

0.5.5 Why is this sub-optimal?

As far back as [**?**], before the popularity of pre-trained word embeddings such as *GloVe* and *Word2Vec*, it was pointed out that using a single token is somewhat crude. It could not possibly encompass the wealth of linguistic information expressed by every OOV word that is encountered; consider that an OOV word can be anything from a spelling mistake to proper noun, such as the name of an entity, and anything in between.

When training an n-gram model, the use of a single $<UNK>$ label for all OOV words will lead to a substantial inconsistency in the frequency of OOV words between training and test datasets [**?**]. This inconsistency is comparable to the possibly counter productive training that is carried out on the singular $<UNK>$ vector across different samples within the scope of sentiment analysis and word embeddings.

In their work, dealing particularly with OOV tokens within the field of Reading Comprehension (RC), [**?**] note a considerable drop in performance when taking this approach in some cases and suggest that a unique OOV token would lack the desired level of detail to correctly generate a correct answer.

0.5.6 Why are classes better for handling OOV words?

It is not necessarily useful to approach the OOV word challenge at the word-level. It is assumed that OOV words are scarce, which substantially limits the occasions for a prospective model to learn any discerning information about that word. Attempting to model clusters of OOV words instead, would benefit each member of the cluster by the accumulated frequency of all members [**?**].

Moreover, within the scope of sentiment analysis, the intuition for clustering words under classes characterized by some particular sentimental value, or a lack thereof; as in the case of registered trademarks, would be evidently beneficial.

0.5.7 What is the trade-off between too many and too few classes?

Too few classes may not possess a sufficiently fine level of detail in their discerning characteristics and cluster together words which are unrelated and subsequently erroneously trained together. This can be seen from the extreme of this case, where only a single token is used, and the issues that have been reported for this approach.

Conversely, if an excessive number of classes are used, this would naturally decrease the amount of OOV words within each class, and consequently the frequency

of words appearing in a particular sample. This hinders a model's ability to learn any distinguishing characteristics of a class. Taken to the extreme, if each class were to contain only a single word, this would effectively render each class as a randomly initialized vector for this word which is rarely encountered, and trained. This undermines the purpose of a word vector, which is to convey as much information about the word as possible.

0.5.8 Why do i think there are benefits to be had from better OOV handling in SA?

Within the scope of a RC task, [**?**], carry out a study to accurately measure the effects that different embeddings and OOV approaches can have on the final result of two benchmark models.

They outline the typical approach to RC problems as initially generating a representation of the source document, possibly through the use of pre-trained word embeddings such as *GloVe* in conjunction with statistical models such as the LSTM [**?**] which may employ an attention mechanism (eg. [**?**]). The result of this process is a contextual representation of the document from which a valid answer can be extracted.

It is worth noting that this process is not dissimilar from the majority of approaches that have been adopted recently within the field of sentiment analysis. Both employ similar techniques and maintain the same characteristic order of events in generating a substantive representation of the source, differing only in the objective and hence the final product to be extracted from that representation. While this is by no means an insignificant difference, on a macro level this can be seen merely as adjusting the variables and parameters that are input to the system, as opposed to the system as a whole.

In their work, [**?**] suggest that there are notable effects on the downstream results of models when comparing the use of different word embeddings, pre-trained or otherwise. Specifically, as an out-of-the-box solution, they recommend the use of GloVe [**?**] 200-dimension pre-trained embeddings. Moreover, for their benchmark RC models, they recommend assigning random unique vectors for OOV tokens at test time, possibly due to the fact that subjects in generated responses are likely to be OOV token and proper nouns.

Based on these findings, the aforementioned similarity in the process of tackling RC tasks and SA tasks, along with the challenges that OOV words pose in the field of SA as previously outlined, the study of word embedding choice and OOV approaches and their effects therein is something that we believe merits further investigation.

0.6 Reproducibility

0.6.1 What is reproducibility and why is it important?

The ability to reproduce experiments is the integral basis upon which all disciplines of science are founded. Within many fields, NLP among them, this typically entails adherence to three integral guidelines, namely, (a) the provision of sufficiently detailed methodologies, (b) the release of operational code-bases and (c) access to the dataset(s) with clear details pertaining to any processing and/or stratification strategies used. These guidelines ensure that results can be easily reproduced, evaluated for generalizability and compared to other methods in the field, thereby fostering growth.

0.6.2 What is the state of reproducibility in the field of Targeted sentiment analysis?

In [?], the authors underscore the significance of reproducibility of approaches as well as the generalizability of the results that are reported, and proceed to argue that adherence to the aforementioned tenets has been lacking in recent years, paying particular attention in their work to the field of targeted sentiment analysis.

The authors also draw attention to the fact that a multitude of studies report results on different datasets which often stem from diverse sources that could be composed of language that is centered around a particular topic. Notable still, these datasets also exhibit consequential statistical differences such as the average length of, and/or amount of targets in, a sentence. Occasionally, studies also carry out particular alterations to existing datasets or adopt a specific strategy for merging one or more datasets (eg. [?]). These factors substantially limit the possibility of effectively comparing the novel approaches as they emerge in the field.

Replication studies such as [?] can remedy this issue by attempting to reproduce the studies in a comparative setting, however they outline challenges in this regard as well. The authors note that a number of approaches in the field fail to outline the precise pre-processing steps they adopted, which may have substantial effects on the downstream performance of a model, and thus make reproducing the results difficult. In some situations they also note model settings mentioned that are not necessarily self-evident, such as a "softmax clipping threshold" [?], which, in attempting to reproduce the study, [?] were forced to ignore as they were unfamiliar with the term.

One key observation that [?] also make with regards to deep learning and neural network based approaches such as [?] is the influence that an initial random seed has on the final performance results, particularly when using smaller word embeddings ([?] as cited in [?]). They mention this as the probable reason for other studies [?],[?] (including their own), not being able to recreate the original results reported in [?]. The authors remark that in situations such as this, when

dealing with models of this nature, it would be well-advised to gauge a model's performance over a number of experiment runs as opposed to one.

0.7 Objectives

Following the principles outlined in [**?**], and motivated by the observations made by [**?**] in the field of reading comprehension and the effect of OOV embedding strategies thereof, the objectives of this study are two-fold.

1. We intend to extend the work carried out by [**?**] in the field of TSA to cover a wider range of studies, including those that employ techniques that have since emerged focusing on attention mechanisms and memory networks. This entails the attempt to reproduce these models based on the detail provided in the original studies and subsequently carrying out a comparative evaluation of these approaches across a wide range of datasets from varying domains using a number of different pre-trained word embeddings.

2. Inspired by the work and findings of [**?**], we shall endeavor to investigate the effect that different OOV embedding strategies and pre-trained word embeddings have on the downstream performance of models with respect to TSA. To our knowledge, at the time of writing, this study will be the first to investigate this issue in detail.

To achieve these goals, we make three principal contributions through this work:

- A publicly accessible framework that provides access to a range of frequently cited datasets that have been used in the field of TSA. This framework shall be used to obtain robust performance metrics, such as macro-f1 scores for all models, which have been hitherto unreported for a subset of the models, as well as other informative measures that shed light into the inner workings of the models implemented, where applicable (such as attention heat-maps).

- A comparative evaluation of models across different domains and datasets, using different pre-trained word embeddings to ascertain the degree to which results obtained are reproducible and generalizable.

- Detailed reports on a series of experiments using different OOV embedding strategies, across all implemented models, the results of which will allow us to deduce the degree to which these affect downstream performance and whether an optimal approach can be found that proves to be generally beneficial.

Finally, the proposed framework shall also serve as groundwork for future experimentation into alternative, more sophisticated, OOV embedding approaches while also providing a means of rapidly carrying out comparative evaluations of TSA models across different datasets and pre-trained word embeddings.