Cross-Domain & In-Domain Sentiment Analysis with Memory-Based Deep Neural Networks *

Gianluca Moro, Andrea Pagliarani, Roberto Pasolini and Claudio Sartori
Department of Computer Science and Engineering (DISI)
University of Bologna
Via Cesare Pavese, 47522 Cesena, Italy
(gianluca.moro, andrea.pagliarani12, roberto.pasolini, claudio.sartori)@unibo.it

Keywords: Sentiment Classification, Transfer Learning, Fine-Tuning, Deep Learning, Big Data, Memory Networks

Abstract: Cross-domain sentiment classifiers aim to predict the polarity, namely the sentiment orientation of target text documents, by reusing a knowledge model learned from a different source domain. Distinct domains are typically heterogeneous in language, so that transfer learning techniques are advisable to support knowledge transfer from source to target. Distributed word representations are able to capture hidden word relationships without supervision, even across domains. Deep neural networks with memory (MemDNN) have recently achieved the state-of-the-art performance in several NLP tasks, including cross-domain sentiment classification of large-scale data. The contribution of this work is the massive experimentations of novel outstanding MemDNN architectures, such as Gated Recurrent Unit (GRU) and Differentiable Neural Computer (DNC) both in cross-domain and in-domain sentiment classification by using the GloVe word embeddings. As far as we know, only GRU neural networks have been applied in cross-domain sentiment classification. Sentiment classifiers based on these deep learning architectures are also assessed from the viewpoint of scalability and accuracy by gradually increasing the training set size, and showing also the effect of fine-tuning, an explicit transfer learning mechanism, on cross-domain tasks. This work shows that MemDNN based classifiers improve the state-of-the-art on Amazon Reviews corpus with reference to document-level cross-domain sentiment classification. On the same corpus, DNC outperforms previous approaches in the analysis of a very large in-domain configuration in both binary and fine-grained document sentiment classification. Finally, DNC achieves accuracy comparable with the state-of-the-art approaches on the Stanford Sentiment Treebank dataset in both binary and fine-grained single-sentence sentiment classification.

1 INTRODUCTION

Sentiment analysis deals with the computational treatment of opinion, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes (a survey is in (Liu and Zhang, 2012)). The task is technically challenging but very useful in practice. For instance, companies always want to know customer opinions about their products.

When an understanding is required of whether a plain text document has a positive, negative or neutral orientation, sentiment classification is involved. This supervised approach learns a model from a labelled training set of documents, then applies it to an unlabelled test set, whose polarity (e.g. positive, negative or neutral orientation) has to be found. The typical approach to sentiment classification assumes that both the training set and the test set deal with the same topic. For example, a model is learnt on a set of book reviews and applied to a distinct set of reviews, but always about books. This modus operandi, known as in-domain sentiment classification, guarantees optimal performance given that documents from the same domain are semantically similar. However this approach is often inapplicable in practice, where documents are mostly unlabelled. Tweets, blogs, fora, comments on social networks could bear opinions, but no information is available on whether they are positive, negative or neutral. Document categorisation by human experts is the only way to deal with such a problem in order to learn an in-domain sentiment classifier, but it is infeasible with large text sets.

*This work was partially supported by the project "Toreador", funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No 688797. Thanks to NVIDIA Corporation for the donated Titan GPU used in this work.

It would be advantageous if a model, once learnt on a source domain, could be used to classify document polarity in a distinct target domain. This approach, known as cross-domain sentiment classification, has become a hot research thread due to its practical implications. The biggest obstacle to learning an effective cross-domain sentiment classifier is the language heterogeneity in documents of different domains. For instance, a book could be described as *interesting* or *boring*, whereas an electrical appliance is more likely to be *working* or *noisy*. In such cases, transfer learning or knowledge transfer techniques can take on the problem, as successfully developed in several research threads, from image recognition to discovery gene biological functions (Domeniconi et al., 2016; Domeniconi et al., 2014b; Domeniconi et al., 2014a). Many transfer learning approaches have been proposed during the years, including the usage of multiple classifiers in (Aue and Gamon, 2005), measures of domain similarity in (Blitzer et al., 2007), feature and document alignment in (Pan et al., 2010; He et al., 2011; Zhang et al., 2015b; Domeniconi et al., 2015b; Domeniconi et al., 2015a; Bollegala et al., 2016), iterative refining of category profiles of documents in (Domeniconi et al., 2014c) and knowledge bases in (Franco-Salvador et al., 2015). They are generally based on dense bag-of-words representation and often require heavy parameter tuning. Despite their good performance with small-scale data (e.g. hundreds or few thousands instances), standard transfer learning approaches do not scale well with the number of features and are not the best choice with large-scale data.

The advent of deep learning has brought a more expressive way to encode text, named distributed representation (aka word vectors), alternative to bag-of-words. Bag-of-words loses the ordering of words and ignores their semantics. Distributed representation solves these problems along with the curse of dimensionality, providing a low-dimensional representation (i.e. 300 features are often enough) wherein words are not mutually exclusive and feature configurations correspond to the variation seen in the observed data. The two main model families for learning word vectors are: global factorization methods, such as latent semantic analysis (LSA) by (Deerwester et al., 1990), and local context window methods, such as the skip-gram and the continuous bag-of-words model by (Mikolov et al., 2013), paragraph vector by (Le and Mikolov, 2014), and others proposed by (Mnih and Kavukcuoglu, 2013; Levy and Goldberg, 2014). Methods from the former family leverage statistical information but perform bad on the word analogy task, whereas those from the latter family are better on the analogy task but inadequately utilise the statistics of corpus since they train on separate local context windows instead of on global co-occurrence counts.

Other than choosing the best text encoding, another aspect that affects sentiment analysis tasks as sentiment classification is how to deal with sequential inputs. This problem impacts on text comprehension and allows the detection of sentiment inversions in phrases or sentences. Recurrent nets are often the best choice for tasks that involve sequential inputs. They process an input sequence one element at a time, maintaining a state vector in their hidden units that implicitly contains information about the history of all past elements of the sequence. Recurrent nets are very powerful, but training them is problematic because the backpropagated gradients either explode or vanish over many time steps, as shown by (Bengio et al., 1994). This makes recurrent neural network unable to learn long dependencies in text. The problem was solved by means of Long Short-Term Memory Network (LSTM) by (Hochreiter and Schmidhuber, 1997), which introduced memory cells to store, load and forget relevant information. Recently, new memory-based neural network schemas have been proposed that achieved the state-of-the-art in many tasks, including machine translation, graph tasks (e.g. graph traversal, shortest path, etc.), and question answering tasks. The rationale is to memorise essential information and use it to handle sequential events and perform complex reasoning on top of them. Sentiment analysis and classification typically require complicated relationships to be inferred, such as the detection of polarity shift and sarcasm. Furthermore, transitive reasoning over multiple sentences is sometimes needed to correctly identify the opinion holder, the target, or the sentiment itself.

The contribution of this work is to investigate with massive experiments to what extent two novel memory-based deep neural network architectures (*MemDNN*) perform in cross-domain and in-domain sentiment classification, which are Gated Recurrent Unit (*GRU*) by (Cho et al., 2014) and Differentiable Neural Computer (*DNC*) by (Graves et al., 2016). We have also combined the two *MemDNN* with the use of Global Vectors (GloVe) proposed by (Pennington et al., 2014) in order to allow them to learn from distributed word representation, now a de facto standard representation in deep learning. As far as we know, only *GRU* has been recently applied by (Dai et al., 2017) to a cross-domain sentiment classification in combination with word embeddings for Chinese corpora. GloVe combines the advantages of the other two major model families in literature for learning word vectors. The unsupervised information ex-

tracted by means of GloVe model is an important first step to align heterogeneous domains. Binary and fine-grained (i.e. multi-class) sentiment classifiers has been constructed for both *MemDNN* architectures. Two benchmark datasets have been used for the experiments: Amazon Reviews dataset [2] for document sentiment classification, and Stanford Sentiment Treebank, introduced by (Socher et al., 2013) for single-sentence sentiment classification. In-domain and cross-domain document-level experiments have been done to assess the variation in performance by the amount of labelled data available for training and validating the model. Results have been compared with those in (Domeniconi et al., 2017), where we developed solutions based on both paragraph vectors, a different text representation method, and other machine learning algorithms. In the former paper, Paragraph Vector by (Le and Mikolov, 2014), despite no explicit transfer learning capability, has been shown to achieve cross-domain accuracy equivalent to a Markov Chain method developed ad-hoc for cross-domain sentiment classification in (Domeniconi et al., 2015b). In the latter paper, the same Markov Chain approach has been outperformed by *GRU* with random feature weights initialisation when large-scale labelled data are available for training and validating the model. To enhance the capability of the *MemDNN* sentiment classifiers in cross-domain tasks, fine-tuning is performed on a small set of labelled target instances. Fine-tuning, along with GloVe word representation and the ability of *MemDNN* in modelling relevant sequential information, aid the inter-domain alignment and bring to outstanding cross-domain document classification results. The *MemDNN* based classifiers have also been employed on very large data sets (e.g. million instances), assessing their document-level performance in an in-domain configuration. Binary and fine-grained experiments have been carried out. The outcome has been compared with several variants of Character-level Convolutional Neural Networks (*CharCNN*) proposed by (Zhang et al., 2015a). *DNC* based classifier outperforms the state-of-the-art in both binary and fine-grained configurations, whereas *GRU* with GloVe feature weights initialisation achieves comparable performance with previous techniques. The experimented *MemDNN* methods can be applied to any text, whatever its length and structure. For this reason, single-sentence sentiment classification has also been performed, using Stanford Sentiment Treebank as the benchmark dataset. The accuracy of *MemDNN* techniques is comparable with state-of-the-art methods in both binary and fine-grained settings.

[2] http://jmcauley.ucsd.edu/data/amazon/

2 RELATED WORK

This work encompasses many research threads, including sentiment classification, cross-domain and transfer learning, and deep learning. Relevant research advances are reviewed in this Section, and other methods are mentioned throughout the paper.

2.1 Sentiment Classification

Sentiment classification consists in labelling a plain text based on its polarity (i.e. sentiment orientation). This task is much more difficult than text classification by topic, because some form of discourse analysis is necessary. (Pang et al., 2002) pointed out that the phenomenon of thwarted expectations narrative is common in documents, where an opinion holder sets up a deliberate contrast to earlier discussion. For instance, *"This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up"*. (Turney, 2002) made a similar point, stating that for reviews the whole is not necessarily the sum of the parts. This is pretty obvious if we observe the previous example. In such a case, humans could easily understand the overall polarity, but it is much less easy for a machine, unless it is able to perform discourse analysis and to detect the polarity shift.

To cope with the complexity of sentiment classification, several methods have been attempted. (Tan et al., 2008) and (Qiu et al., 2009) employed a dictionary containing commonly used words in expressing sentiment to label a portion of informative examples from a given domain, in order to reduce the labelling effort and to use the labelled documents as a training set for a supervised classifier. (Melville et al., 2009) exploited lexical information about associations between words and classes, and refined them for specific domains by means of training examples to enhance accuracy. Other works by (Deng et al., 2014; Wu and Gu, 2014; Domeniconi et al., 2015c) proposed term weighting schemes to foster sentiment classification.

Cross-domain comes into play when the target domain lacks (or has few) labelled data for training a classifier with supervision. Transfer learning techniques are generally required to bridge the semantic gap due to language heterogeneity across domains. Two transfer learning modes have been identified by (Pan and Yang, 2010), namely, *instance transfer* and *feature representation transfer*. In order to bridge the

inter-domain gap, the former adapts source instances to the target domain, whereas the latter maps source and target features into a different space. (Aue and Gamon, 2005) made some attempts to customize a classifier to a new target domain: training on a mixture of labelled data from other domains where such data is available, possibly considering just the features observed in target domain; using multiple classifiers trained on labelled data from diverse domains; including a small amount of labelled data from the target. (Blitzer et al., 2007) discovered a domain similarity measure that fosters domain adaptation. (Pan et al., 2010) introduced a spectral feature alignment technique, where domain independent terms helps aligning domain specific terms into the same clusters. These clusters form a latent space that improves the classification of the target domain. Apart from this, other algorithms have been proposed in (Zhang et al., 2015b; Domeniconi et al., 2015b; Domeniconi et al., 2015a) to transfer the polarity of features from the source domain to the target domain by using domain independent features as a bridge. (He et al., 2011) modified the topic-word Dirichlet priors and extended the *joint sentiment-topic* model by adding prior words sentiment. Polarity-bearing topics have been used to perform feature and document expansion so as to align domains. (Bollegala et al., 2013) suggested the adoption of a thesaurus containing labelled data from the source domain and unlabelled data from both source and target domains. (Bollegala et al., 2016) modelled cross-domain sentiment classification as embedding learning, and discovered that a good objective function should capture geometric properties in the unlabelled documents of both source and target domains. These unsupervised properties are even more important than considering common features that occur in both domains and than setting label constraints to the source domain documents.

2.2 Deep Learning

The advent of deep learning has dramatically improved the state-of-the-art in several research areas, such as speech processing and recognition, visual object detection, video, audio, and natural language processing, and many other domains like drug discovery and genomics, as pointed out by (LeCun et al., 2015). The first issue to face when analysing a plain text is how to deal with sequential data. This problem is even more essential to detect sentiment orientation, because of the presence of sarcasm, negations, and the phenomenon of thwarted expectations narrative. Bag-of-words text representation, where the presence (or the frequency) of terms into documents is encoded in a term-document matrix, is intrinsically unable to handle sequential inputs. Word ordering is lost and word semantics is ignored, since context is not taken into account. Another big issue of the bag-of-words model is dimensionality, because each term is a feature of the model, resulting in very sparse term-document matrices. Feature selection techniques attenuate the problem and let data be processed, but relevant information can be lost during this process.

Alternative to the bag-of-words model are distributed text representations. Words are mapped into low-dimensional vector spaces, where features, called word vectors, capture most of the variation observed in data. A feature in the newer space incorporates the characteristics of several features in the original space. (Mikolov et al., 2013) introduced the continuous bag-of-words (CBOW) model, a local context window method derived from the neural network language model by (Bengio et al., 2003). In CBOW, a projection layer is shared among words so that their vectors get projected (e.g. averaged) into the same position. The model is trained by building a log-linear classifier with k future and k history words as input, where the training criterion is to correctly predict the current word. In the same work, the skip-gram model is also proposed, where the current word is used as input to a log-linear classifier with continuous projection layer to predict words within a certain range before and after the current word itself. Following the same idea of word vectors, (Le and Mikolov, 2014) proposed an approach to learn paragraph vectors. Every paragraph vector is mapped into a unique vector, then averaged or concatenated to word vectors to predict the next word in a given window size (i.e. context). In spite of capturing semantic and syntactic regularities, local context window methods for distributed word representation typically fail in modelling global statistics and properties. (Pennington et al., 2014) advanced a global log-bilinear regression model to solve this lack. Their GloVe model utilises the benefits of count-based methods like LSA by (Deerwester et al., 1990), while simultaneously capturing the meaningful linear substructures prevalent in local context window methods.

Along with the distributed word representation models, several deep learning architectures have been proposed that brought to a dramatic improvement in sentiment classification. (Dos Santos and Gatti, 2014) proposed a deep convolutional neural network that jointly uses character-level, word-level and sentence-level representations to perform sentiment analysis of short texts. (Socher et al., 2013) introduced Recursive Neural Tensor Networks (*RecNTN*) for single-sentence sentiment classification. Its recursive struc-

ture makes *RecNTN* able to capture polarity shifts in sentences. The experiments have been carried out on Stanford Sentiment Treebank, which became a benchmark for single-sentence sentiment classification. It turned out that *RecNTN* improves the state-of-the-art in both binary and fine-grained configurations. *RecNTN* has been outperformed by the Dynamic Memory Network (*DMN*) by (Kumar et al., 2016), which naturally captures position and temporality by processing input sequences and questions, forming episodic memories, and generating relevant answers. The memory and input modules of the original technique have been improved later by (Xiong et al., 2016). (Tang et al., 2015) introduced Gated Recurrent Neural Networks to learn vector-based document representation, showing that the underlying model outperforms the standard Recurrent Neural Networks in document modelling for sentiment classification. (Zhang and LeCun, 2015) applied temporal convolutional networks to large-scale data sets, showing that they can perform well without the knowledge of words or any other syntactic or semantic structures.

Despite the success of deep nets, few work has been done on transfer learning and cross-domain sentiment classification so far. The Stacked Denoising Autoencoder, introduced in (Vincent et al., 2010), was used by (Glorot et al., 2011) to extract domain-independent features without supervision that act as a bridge between heterogeneous domains. In (Domeniconi et al., 2017), we showed that labelled data from multiple domains encoded by means of paragraph vectors help transfer learning and cross-domain sentiment classification.

3 DEEP LEARNING ADVANCES

This section describes the main features of the deep learning advances combined in this work to break through cross-domain sentiment classification.

3.1 Gated Recurrent Unit

Gated Recurrent Unit (*GRU*), proposed by (Cho et al., 2014), is an evolution of Long Short-Term Memory (*LSTM*), a neural network architecture provided with a memory mechanism that allows storing and retaining information through long time sequences. *GRU* adds a mechanism that makes each recurrent unit adaptively able to capture dependencies of different time scales. While *LSTM* is composed of three gates (i.e. input, output, and forget), *GRU* only has two gates, such as update and reset (figure 1). The update gate rules the unit activation, by deciding how much information will be moved from the previous hidden state to the current one. Any information in the hidden state that becomes irrelevant later on is dropped via the reset gate. As each hidden unit has separate reset and update gates, it will learn to capture dependencies over different time scales.

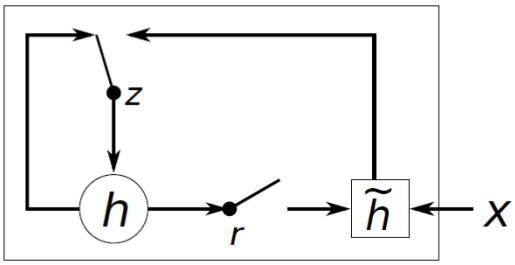

Figure 1: A schematic representation of GRU.

GRU with randomly initialised feature weights has shown promising results in cross-domain sentiment classification with large-scale data. When enough training instances are available, the alignment of heterogeneous domains is achieved thanks to memory units, which are automatically able to capture and preserve domain-independent information, despite no explicit transfer learning mechanism.

3.2 Differentiable Neural Computer

Differentiable Neural Computer (*DNC*), introduced by (Graves et al., 2016) as the evolution of Neural Turing Machines (*NTM*) by (Graves et al., 2014), is one of the most innovative *MemDNN* techniques. Differently from previous *MemDNN* architectures (e.g. *GRU*), where the memory mechanism was internal to the network, *DNC* uses an external memory to represent and manipulate complex data structures. The neural network can selectively address the external memory, both to read from and write to it, allowing iterative modification of memory content. This makes *DNC* able to learn complex tasks from data, such as finding the shortest path or inferring the missing links in graphs, and answering synthetic questions designed to emulate reasoning in natural language. Figure 2 shows the basic behaviour of a *DNC*. It uses differentiable attention mechanisms to define weightings, which represent the degree to which each memory location is involved in a read or write operation. The functional units that determine and apply the weightings are called read and write heads.

In the original work, *DNC* has only been applied to small-scale tasks. However, (Graves et al., 2016) pointed out that *DNC* should be able to seamlessly

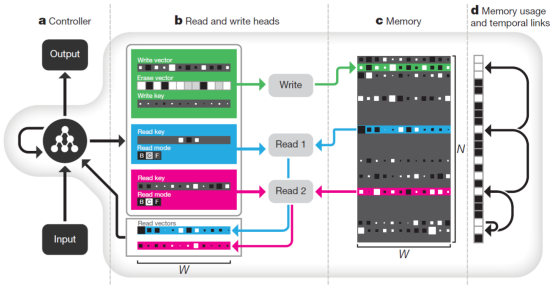

Figure 2: A schematic representation of DNC.

acquire knowledge and take advantage of exposure to large data sources. This consideration, along with the ability of memory mechanisms to capture inter-domain relationships, makes *DNC* suitable for cross-domain sentiment classification.

3.3 GloVe Word Representation

Global Vectors (GloVe) is a log-bilinear regression model that have been proposed by (Pennington et al., 2014) to learn distributed word representation. Alike other methods for learning vector space representation of words, GloVe is able to capture fine-grained syntactic and semantic regularities in an unsupervised fashion, just using vector arithmetic, and solves the data sparsity problem of dense bag-of-words models. GloVe combines the advantages of global matrix factorization and local context window methods: as the former, it efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix; as the latter, it achieves great performance on word analogy, similarity and named entity recognition tasks. The unsupervised information extracted by means of distributed word representation fosters the alignment of heterogeneous domains ; for this reason, we argue that GloVe can be promising to initialise the feature weights that *MemDNN* architectures will use.

4 DATASETS

In this Section the benchmark datasets for document-level and single-sentence classification respectively will be introduced. Amazon Reviews corpus[3] has been used for the former task, whereas Stanford Sentiment Treebank[4] for the latter. Both are widely used benchmarks for sentiment analysis.

[3]http://jmcauley.ucsd.edu/data/amazon/

[4]https://nlp.stanford.edu/sentiment/code.html

4.1 Amazon Reviews Corpus

Amazon Reviews corpus is a collection of Amazon reviews about different domains. Each domain contains a list of English reviews, which include both the plain text and a score from 1 (i.e. very negative) to 5 (i.e. very positive). In binary sentiment classification, reviews with rating > 3 have been considered as positive, reviews with rating < 3 as negative, while reviews with rating $= 3$ have been discarded as they are ambiguous and could express a neutral sentiment orientation. On the other hand, all the 5 classes have been taken into account in the fine-grained setting. Data from 4 domains have been used for the experiments: Books (B), Movies (M), Electronics (E) and Clothing-Shoes-Jewelry (J) have been chosen for a matter of comparison with the state of the art.

4.2 Stanford Sentiment Treebank

Stanford Sentiment Treebank (SST) is a dataset of labelled sentences that was introduced by (Socher et al., 2013). SST is built on a corpus of movie review excerpts, composed of 11,855 sentences, half of which are positive and half negative. The sentences are parsed with the Stanford parser by (Klein and Manning, 2003) into 215,154 syntactically plausible phrases. Each phrase is annotated by 3 human experts into 5 possible categories, namely negative, somewhat negative, neutral, somewhat positive and positive. Similarly to Amazon Reviews corpus, neutral phrases are discarded in binary classification.

5 EXPERIMENTS AND RESULTS

This Section illustrates the experiments that have been performed. The first assesses to what extent the amount of labelled data available for training the model affects its performance in both in-domain and cross-domain document sentiment classification. Then the impact of fine-tuning on cross-domain is evaluated, with appropriate comparison with the state of the art. In the third experiment, in-domain document sentiment classification is performed on large-scale data, in order to evaluate the scalability of *MemDNN* techniques and their potential feasibility in big data scenarios. The last trial assesses whether *MemDNNs* can be successfully applied to single-sentence sentiment classification.

Accuracy of the classifier (i.e. the percentage of correctly classified instances) has been measured for each single test, averaging results on 10 randomly chosen training-test partitions to reduce the variance,

(i.e. the sensitivity to small variations in the training set), but always keeping the classes balanced.

5.1 The Impact of Training Data

The first experiment checks to what extent the amount of labelled training data affects *MemDNN* performance. Naïve Bayes (*NB*), Markov Chain (*MC*), Paragraph Vector (*PV*) and Gated Recurrent Unit with randomly initialised feature weights (GRU_{rand}) have already been taken into account in (Domeniconi et al., 2017). For a matter of comparison, source-target partitions of three different orders of magnitude have been tested, preserving 80%-20% as the source-target ratio, and balancing positive and negative examples. The small-scale data set has 1,600 labelled instances as the training set and 400 unlabelled instances as the test set; the medium-scale 16,000 and 4,000; and the large-scale 80,000 and 20,000 respectively.

Figure 3 shows the in-domain performance of the various techniques, averaged on the 4 domains considered (detailed results have not been reported due to space reason). As pointed out by (Domeniconi et al., 2017), deep learning approaches usually do not perform well when few training data are available. That is the reason why *MC* outperforms the proposed *MemDNN* techniques with small-scale data. However, *GRU* and *DNC* outperform the other approaches. *GRU* with feature weights initialised by GloVe achieves a higher accuracy with respect to GRU_{rand} whose features have been initialised with random weights. Increasing the amount of labelled training data, *DNC* obtains astonishing performance. Its accuracy is 90.08% with medium-scale data, meaning that 16,000 training examples are enough for the memory mechanism of *DNC* to capture relevant information. The same does not hold for *GRU*, whose performance does not increase considering medium-scale data. However, in opposition to their trial, *GRU* already achieves comparable performance with *MC* in the medium-sized data set. Considering large-scale data, the accuracy of *DNC* continues to grow, reaching 91.24%. This outcome makes it interesting to evaluate to what extent *DNC* performance can increase. For this purpose, an in-domain test with a huge dataset will be shown later in 5.3. Finally, it may be noted that *GRU* performance improves as well. A reasonable explanation is that the memory mechanism of *GRU* is automatically able to decide which information is relevant to classification, if trained with a large amount of data.

The cross-domain evaluation of the same techniques can be seen in figure 4. The plot displays accuracy averaged on each of the 12 source-target con-

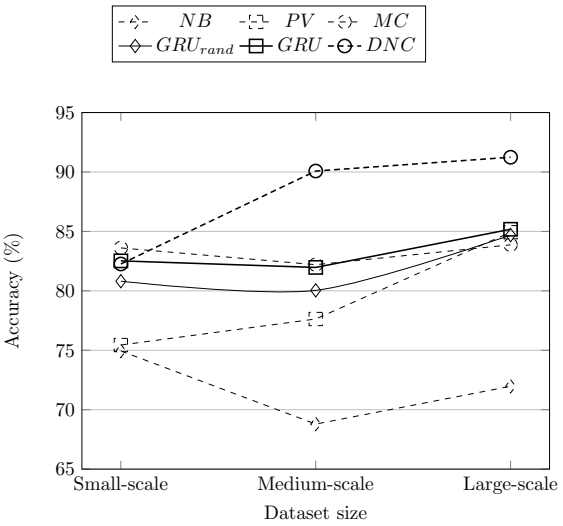

Figure 3: Average in-domain accuracy over the 4 domains.

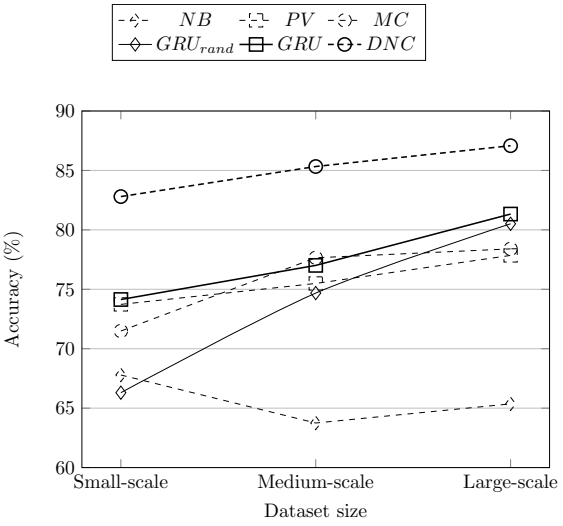

Figure 4: Average cross-domain accuracy over the 12 source-target combinations of the 4 domains.

figurations of the 4 domains considered. The first outcome that catches the eye is that *DNC* dramatically outperforms all the other techniques regardless of the dataset size. It is remarkable that *DNC* exceeds by more than 9% the accuracy of *MC*, which is a non-deep method that was specifically developed by (Domeniconi et al., 2015b) to accomplish both transfer learning and sentiment classification. The reason of this outcome resides in several combined factors that lead to semantic comprehension of text. The first factor is the usage of distributed representation to encode text. In particular, we used GloVe for word representation of both *GRU* and *DNC*. As pointed out by (Pennington et al., 2014), GloVe combines the advantages of the other two model families in literature

for learning word vectors, namely factorization methods and local context window methods. This means that GloVe also inherits the benefits of *PV*, which is able to discover hidden relationships between semantically similar words. The unsupervised information extracted by GloVe aids the alignment of heterogeneous domains. The second factor is the memory mechanism of *DNC*. Once enough training data are available, *MemDNN* architectures are automatically able to capture domain-independent information and preserve it in memory. The third factor are deep neural networks. In particular, *DNC* is one of the most powerful mechanisms to emulate reasoning and inference problems in natural language. The combined effect of these three factors led to a dramatic improvement of the state of the art in cross-domain sentiment classification. *DNC* turns out to be 9% more accurate than *MC*. Comparing in-domain and cross-domain results, it could be noted that the accuracy of *DNC* is perfectly aligned by looking at small-scale data, whereas cross-domain performance is slightly worse by increasing the dataset size. Apart from the astonishing performance of *DNC*, careful readers can note the behaviours of *GRU* and GRU_{rand} respectively, which probably are even more interesting. As expected, GloVe initialisation of feature weights leads to a substantial increasing of accuracy with small-scale data, which jumps from 66.30% of GRU_{rand} to 74.14% of *GRU*. Comparing in-domain and cross-domain experiments, we can see the combined effect of GloVe distributed word representation and *GRU* memory mechanism. The former plays a key role to align heterogeneous domains when few labelled data are available as the training set. The latter is automatically able to extract relevant inter-domain concepts as the amount of labelled training data increases.

5.2 Fine-Tuning of MemDNNs

The second experiment aims to assess whether fine-tuning affects *MemDNN* performance. Fine-tuning is the practice of using a labelled sample of target instances to refine a model previously learnt on the source domain. The sample is usually small (e.g. hundreds instances) for two main reasons. On the one hand, if a large set of labelled instances was available, it would be advisable to learn an in-domain sentiment classifier rather than a cross-domain one. On the other hand, if a large set of labelled instances was not available, the only alternative would be to let a team of human experts pre-classify some instances. Manual categorisation becomes infeasible when many instances are required to be labelled. Therefore, fine-tuning on a small sample of labelled target instances is generally a good trade-off between its cost and the expected improvement of performance. To further investigate the performance, we have experimented fine-tuning by using 250 and 500 examples respectively.

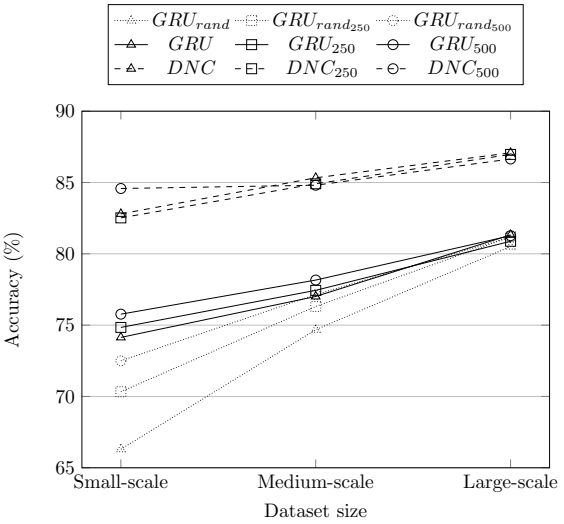

Figure 5: Average cross-domain with fine-tuning accuracy over the 12 source-target combinations of the 4 domains. The subscripts 250 and 500 represent the number of labelled target instances utilised for fine-funing.

Figure 5 shows the effect of fine-tuning on *MemDNN* architectures. *GRU* takes a slight advantage of fine-tuning. With reference to small-scale data, accuracy increases from 74.14% to 74.84% using 250 target instances to 75.77% using 500 target instances. It deserves to be noted that *GRU* performance is more affected by GloVe feature weights initialisation than by fine-tuning. In fact, *GRU* without fine-tuning outperforms $GRU_{rand_{500}}$. As dataset size increases, the contribution of fine-tuning diminishes, until eventually vanishing with large-scale data. A different behaviour is observed for *DNC*. Fine-tuning only impacts on accuracy when performed on 500 target instances with reference to the small-scale dataset. It is quite obvious that, when few training data are available, even a small sample can considerably affect performance. *GRU* is a clear proof of this behaviour. However, DNC_{250} does not lead to a performance improvement. The reason is that *DNC* is a very robust technique, almost unaffected by noise. The attention mechanism to address the external memory makes *DNC* less sensitive to noise than *GRU*, whose memory units are internal to the network. As a consequence, *DNC* is less prone than *GRU* to altering memory content. In other words, it is unlikely that *DNC* stores irrelevant information in memory. 250 target instances are not relevant enough for *DNC* and are considered as noise by the network. The same con-

siderations apply to experiments with medium-scale and large-scale data, where both 250 and 500 target instances do not affect performance.

5.3 Large-Scale Classification

The third experimentation is an in-domain sentiment classification task with a very large data set. This trial let us assess to what extent *MemDNN* architectures with GloVe feature weights initialisation are suitable as sentiment classifiers in big data scenarios. Moreover, *MemDNN* architectures can be combined with many other sentiment classification techniques, in particular several variants of Character-level Convolutional Neural networks (*CharCNN*), proposed by (Zhang and LeCun, 2015) and empirically explored in (Zhang et al., 2015a). For result comparison purposes, two very large data sets have been constructed. The former deals with binary in-domain sentiment classification, where the goal is to distinguish positive from negative instances. The latter aims to predict the full score assigned to instances (i.e. from 1 to 5). So it is a fine-grained in-domain sentiment classification task. The binary dataset contains $1,800,000$ training samples and $200,000$ testing samples for each polarity sentiment. The fine-grained contains $600,000$ training samples and $130,000$ testing samples for each of the five classes. In both datasets, samples have been taken in equal proportion from the 4 domains considered. Differently from the previous experiments, review title has also been considered, together with review content.

Apart from the several variants of *CharCNN*, results have also been compared with other methods, including Long Short-Term Memory networks (*LSTM*), Bag-of-means by (Lev et al., 2015) and some Bag-of-words (*BoW*) based configurations. Careful readers can find further details on these methods along with their parameters in (Zhang et al., 2015a). Table 1 shows the accuracy of *MemDNN* methods and the state-of-the-art techniques. *GRU* achieves comparable performance with the other methods. In particular, it is slightly more accurate than *LSTM*. This is not surprising, since *GRU* is an evolution of *LSTM*, but both have a built-in memory mechanism. On the other hand, *DNC* outcome is astonishing. It outperforms all the other techniques with reference to both binary and fine-grained datasets. Fine-grained accuracy is almost 2% higher than the previous methods. This difference in accuracy is significant in a multinomial classification problem, where predicting the correct class is challenging. To the best of our knowledge, it is the first time that a method achieves accuracy higher than 60% on Amazon Reviews corpus in fine-grained sentiment classification.

Table 1: In-domain accuracy on very large datasets constructed from Amazon Reviews corpus. Binary and fine-grained refer to 2-class and 5-class in-domain sentiment classification respectively. *CharCNN* variants are prefixed with *Lg.* or *Sm.*.

Model	Binary	Fine-grained
BoW	90.40%	54.64%
BoW Tf-Idf	91.00%	55.26%
n-grams	92.02%	54.27%
n-grams Tf-Idf	91.54%	52.44%
Bag-of-means	81.61%	44.13%
LSTM	93.90%	59.43%
Lg. w2v Conv	94.12%	55.60%
Sm. w2v Conv	94.00%	57.41%
Lg. w2v Conv. Th.	94.20%	56.25%
Sm. w2v Conv. Th.	94.37%	57.50%
Lg. Lk. Conv	94.16%	54.05%
Sm. Lk. Conv	94.15%	56.34%
Lg. Lk. Conv. Th.	94.48%	57.61%
Sm. Lk. Conv. Th.	94.49%	56.81%
Lg. Full. Conv	94.22%	59.11%
Sm. Full. Conv	94.22%	59.12%
Lg. Full. Conv. Th.	94.49%	59.46%
Sm. Full. Conv. Th.	94.34%	59.47%
Lg. Conv	94.49%	58.69%
Sm. Conv	94.50%	59.47%
Lg. Conv. Th.	95.07%	59.55%
Sm. Conv. Th.	94.33%	59.57%
GRU	94.07%	59.55%
DNC	**95.51**%	**61.45**%

5.4 Single-Sentence Classification

While the previous experiments deal with document sentiment classification, the last one focuses on single-sentence sentiment classification. The benchmark dataset used is Stanford Sentiment Treebank (SST). According to the work by (Socher et al., 2013), $8,544$ sentences are used as the training set, $1,101$ as the validation set, and $2,210$ as the test set. Plenty of techniques have been applied to SST in the last few years. (Socher et al., 2013) presented Recursive Neural Tensor Networks (*RecNTN*) in the same work they introduced SST, and compared their algorithm on SST with Naïve Bayes with unigram features (*NB*), Naïve Bayes with unigram and bigram features (*BiNB*), Support Vector Machine with unigram and bigram features (*SVM*), Recursive Neural Networks (*RNN*) by (Socher et al., 2011) and Matrix-Vector RNN (*MV-RNN*) by (Socher et al., 2012). (Kalchbrenner et al., 2014) proposed Dynamic Convolutional Neural Network (*DCNN*), comparing its performance on SST with Max Time-Delay Neural Networks (*MaxTDNN*) by (Collobert and Weston, 2008), and a Neural Bag-of-Words (*NBoW*) model. (Dos Santos and Gatti, 2014) introduced Character to Sentence Convolutional Neural Network (*CharSCNN*) and applied it to SST. A variant of *CharSCNN* has been trained by using word embeddings only (*SCNN*). Other two variants of the previous, referred as *CharSCNN ph.*

and *SCNN ph.*, have been trained by exploiting also phrases representation in addition to sentence representation. (Kim, 2014) experimented some variants of Convolutional Neural Networks (*CNN-rand*, *CNN-static*, *CNN-non-static*, *CNN-multichannel*) on SST. (Le and Mikolov, 2014) applied to SST logistic regression on top of their Paragraph Vector *PV* for distributed word representation. Finally, Multiplicative Recurrent Neural Network (*DRNN*) by (Irsoy and Cardie, 2014), Constituency Tree-LSTM (*CT-LSTM*) by (Tai et al., 2015), and Dynamic Memory Network (*DMN*) by (Kumar et al., 2016) have also been applied to SST.

Table 2: Accuracy achieved by the compared methods on SST. Binary and fine-grained refer to 2-class and 5-class in-domain sentiment classification respectively.

Model	Binary	Fine-grained
NB	81.80%	41.00%
BiNB	83.10%	41.90%
SVM	79.40%	40.70%
RecNTN	85.40%	45.70%
Max-TDNN	77.10%	37.40%
NBoW	80.50%	42.40%
DCNN	86.80%	48.50%
RNN	82.40%	43.20%
MV-RNN	82.90%	44.40%
SCNN	82.00%	43.50%
CharSCNN	82.30%	43.50%
SCNNph.	85.50%	48.30%
CharSCNNph.	85.70%	48.30%
CNN-rand	82.70%	45.00%
CNN-static	86.80%	45.50%
CNN-non-static	87.20%	48.00%
CNN-multichannel	88.10%	47.40%
PV	87.80%	48.70%
DRNN	86.60%	49.80%
CT-LSTM	88.00%	51.00%
DMN	**88.60%**	**52.10%**
GRU	84.13%	45.89%
DNC	85.22%	46.78%

Table 2 shows the comparison between the *MemDNN* architectures and the mentioned methods. *GRU* and *DNC* achieve comparable performance in both binary and fine-grained configurations. The accuracy of *DNC* is just about 1% higher than the accuracy of *GRU*. They perform similarly to most of the other techniques, but are not definitely the best methods for single-sentence sentiment classification. This is probably due to the absence of a specific mechanism to take sentence syntax into account, and to the small amount of training data, which is an obstacle to *GRU* and *DNC* performance. Just look at the in-domain experiment on Amazon Reviews 3, where they have been outperformed by Markov Chain with small-scale data. Somebody might argue that SST have $8,544$ instances, but we should not forget that they are single-sentences, not whole and usually longer reviews (i.e. documents) as in the Amazon dataset. The best algorithm turns out to be *DMN*, which performs better than all the other techniques in both binary and fine-grained configurations. This is not surprising, since *DMN* includes a memory mechanism to store and preserve relevant information through time and has also been proved to work well with single-sentences in (Kumar et al., 2016).

6 CONCLUSIONS

This work has investigated with massive experiments to what extent novel memory-based neural networks (*MemDNN*) perform in cross-domain and in-domain sentiment classifications. We have combined the advances of *MemDNN* together with word embeddings, a de facto standard in deep learning, along with fine-tuning on target instances to investigate whether they are able to outperform ad-hoc cross-domain solutions. Among the deep memory-based methods, we experimented Differentiable Neural Computer and Gated Recurrent Unit. The former is one of the most innovative deep learning techniques. Its ability to address and manage an external memory makes *DNC* able to emulate reasoning and inference problems in natural language. The latter is a different kind of *MemDNN*, since its memory mechanism is part of the network structure. GloVe distributed word representation has been used in combination with both *MemDNN* architectures.

Experiments on Amazon Reviews corpus show that *DNC* with GloVe word representation dramatically outperforms state-of-the-art techniques for cross-domain sentiment classification. Transfer learning from the source to the target domain is supported by distributed word representation with small-scale datasets, as proved by the comparison between *GRU* and GRU_{rand}, and by memory mechanisms as the dataset size increases. *MemDNN* techniques take advantage of large-scale data to align heterogeneous domains. Fine-tuning on a small sample of target instances is more useful to *GRU* than *DNC*, as the latter is more robust and less sensitive to noise. Both techniques have been compared with state-of-the-art methods on two very large datasets, built on the same Amazon Reviews corpus, for in-domain document sentiment classification. *DNC* with GloVe feature weights achieves new state-of-the-art performance both in binary and fine-grained classification tasks. Finally, *DNC* and *GRU* achieve comparable performance with many techniques in single-sentence in-domain sentiment classification on Stanford Sentiment Treebank. Small-scale training data and the absence of a mechanism to deal with sentence syntax are probably the reasons that prevent *DNC* from reaching the state-of-the-art performance.

REFERENCES

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *RANLP*, volume 1, pages 2–1.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *45th ACL meeting*, pages 440–447.

Bollegala, D., Mu, T., and Goulermas, J. Y. (2016). Cross-domain sentiment classification using sentiment sensitive embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):398–410.

Bollegala, D., Weir, D., and Carroll, J. (2013). Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE transactions on knowledge and data engineering*, 25(8):1719–1731.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *the 25th ICML*, pages 160–167. ACM.

Dai, M., Huang, S., Zhong, J., Yang, C., and Yang, S. (2017). Influence of noise on transfer learning in chinese sentiment classification using gru. In *ICNC-FSKD*, pages 1844–1849. IEEE.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Deng, Z.-H., Luo, K.-H., and Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7):3506–3513.

Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2014a). Discovering new gene functionalities from random perturbations of known gene ontological annotations. In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 107–116. SciTePress.

Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2014b). Random perturbations of term weighted gene ontology annotations for discovering gene unknown functionalities. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers*, volume 553 of *Communications in Computer and Information Science*, pages 181–197. Springer.

Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2016). Cross-organism learning method to discover new gene functionalities. *Computer Methods and Programs in Biomedicine*, 126:20–34.

Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2015a). Cross-domain sentiment classification via polarity-driven state transitions in a markov model. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2015, Lisbon, Portugal, 2015, Revised Selected Papers*, volume 631 of *Communications in Computer and Information Science*, pages 118–138. Springer.

Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2015b). Markov chain based method for in-domain and cross-domain sentiment classification. In *KDIR 2015 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 1, Lisbon, Portugal, 2015*, pages 127–137. SciTePress.

Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2017). On deep learning in cross-domain sentiment classification. In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - (Volume 1), Funchal, Madeira, Portugal, November 1-3, 2017.*, pages 50–60. SciTePress.

Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2014c). Iterative refining of category profiles for nearest centroid cross-domain text classification. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - IC3K 2014, Rome, Italy, 2014, Revised Selected Papers*, volume 553 of *Communications in Computer and Information Science*, pages 50–67. Springer.

Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2015c). A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf.idf. In *Data Management Technologies and Applications - 4th International Conference, DATA 2015, Colmar, France, 2015, Revised Selected Papers*, volume 584 of *Communications in Computer and Information Science*, pages 39–58. Springer.

Dos Santos, C. N. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78.

Franco-Salvador, M., Cruz, F. L., Troyano, J. A., and Rosso, P. (2015). Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46–56.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *the 28th ICML*, pages 513–520.

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G.,

Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

He, Y., Lin, C., and Alani, H. (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *49th ACL Meeting: Human Language Technologies-Volume 1*, pages 123–131.

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

Irsoy, O. and Cardie, C. (2014). Modeling compositionality with multiplicative recurrent neural networks. *arXiv preprint arXiv:1412.6577*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st ACL meeting*.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. (2016). Ask me anything: Dynami memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *31st ICML*, pages 1188–1196.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lev, G., Klein, B., and Wolf, L. (2015). In defense of word embedding for generic text representation. In *NLDB*, pages 35–50. Springer.

Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *15th ACM SIGKDD*, pages 1275–1284. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *19th international conference on World wide web*, pages 751–760. ACM.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP, Volume 10*, pages 79–86. ACL.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Qiu, L., Zhang, W., Hu, C., and Zhao, K. (2009). Selc: a self-supervised model for sentiment classification. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 929–936.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *EMNLP-CoNLL*, pages 1201–1211. ACL.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP*, pages 151–161. ACL.

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 *EMNLP*, pages 1631–1642.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Tan, S., Wang, Y., and Cheng, X. (2008). Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *31st ACM SIGIR conference on Research and development in information retrieval*, pages 743–744. ACM.

Tang, D., Qin, B., and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *40th ACL meeting*, pages 417–424.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408.

Wu, H. and Gu, X. (2014). Reducing over-weighting in supervised term weighting for sentiment analysis. In *COLING*, pages 1322–1330.

Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *ICML*, pages 2397–2406.

Zhang, X. and LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Zhang, X., Zhao, J., and LeCun, Y. (2015a). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhang, Y., Hu, X., Li, P., Li, L., and Wu, X. (2015b). Cross-domain sentiment classification-feature divergence, polarity divergence or both? *Pattern recognition letters*, 65:44–50.