Capsule layer and attention layer augmentation analysis over convolutional methods for review categorization in Amazon dataset

Sharat Sachin, Abha Tripathi, Navya Mahajan, Shivani Aggarwal, Preeti Nagrath
Bharati Vidyapeeth's College of Engineering, New Delhi

Abstract. The world as we know it today is becoming more and more technologically advanced. E-commerce is one of the most widely adopted of these advancements, now seeing penetration all around the world. Now people sitting in the comfort of their homes can order items and have them at their fingertips in days, and some cases, hours. For these people relying on the online items, one of the most important metrics is the reviews received by that item. To obtain a balanced view of the item, the customer may need to see the positive as well as negative reviews for that item. Modern day algorithms and models can be made to decide the reviews seen by the user so he can judge an item fairly. We have attempted to make models that can judge the sentiment of the user writing the reviews from the text of the review, classifying it into positive, negative or neutral. In this work, we have performed a comparison of the effect of convolutional layers, attention layers and capsule network layers on base models for GRU, Bi-GRU, LSTM and Bi-LSTM on an Amazon review dataset used for classification.

Keywords: Deep learning, capsule network, sentiment analysis, convolutional neural networks, attention, recurrent neural networks, long short-term memory, gated recurrent units.

1 Introduction

Sentiment analysis, denoted as idea or sentiment mining, is the art of extracting emotion from blocks of text, with no help from human actors so as to categorize them according to the polarity of the sentiment. We attempt to glean positive or negative intent from text, and try to also categorize cases where neither of the positive and negative overwhelms the other and the overall intent can thus be classified as neutral. Examining and processing the full text, sentiment analysis tries to extract sentiment information from it.

With the rate at which internet usage has changed for people in recent times, the different manners in which production and dissemination of content occurs become more and more diverse, especially considering different forms of text and reviews by the users [1]. A significant part of this contains phrases from which humans can glean emotional awareness, yet this poses significant challenges for a learning system. A business can gain a significant advantage by keeping abreast of the consensus relating to its products online, on interfaces such as review sites and product forums. This can also influence decisions made concerning customers [2]. It helps companies to analyse and improve their brand image by getting access to clean and meaningful data, and focus their attention on improving aspects of the business that have low customer satisfaction [3].

RNN's are models that are used in various fields of application dealing with sequenced data, however they are limited by the exploding and vanishing gradient problems [4]. The vanishing gradient problem was successfully resolved by Hochreiter & Schmidhuber [5], using LSTM architecture. The GRU (which was developed by Kyunghyun Cho et al in 2014 [6]) performs all the functions performed by the LSTM unit, but it accomplishes these functions without employing any extra memory units. The bidirectional variants of these models are designed with a specific motivation in mind - that the context of a word is defined by what precedes that word as well as what follows it. Thus, these models act with data from both backwards and forwards in time to make a decision. The CapsNet (Capsule neural

network) is a technique that attempts to add structures that better mimic relationships that contain a hierarchical collection of objects. This consists of capsules which are a group of neurons that activate uniquely for different types of characteristics of a different entity, and we reuse output from these capsules to form more balanced representations for higher ranked capsules [7].

Models that are based on deep learning have exhibited great promise in determining the polarity of text in recent times, with a variety of research focusing on text classification by polarity [8]. One reason for this is that they are able to learn the intent of the writer from training data without tiresome feature engineering. Sentiment mining is a very important field of natural language processing and has been studied in various fields [9]. In this publication, the main focus is on comparing the results of using baseline unidirectional and bidirectional LSTM and GRU to ones augmented using a convolutional layer and self-attention layer. We also use the state-of-the-art capsNet layer which is augmented to the various base models and then compare the accuracy achieved with them.

Descriptions of the convolutional neural networks, gated recurrent architectures, attention mechanism and capsule network are discussed in the second section where we have worked on this review. The next section consists of the literature review, that consists of descriptions of the previous research performed in various fields of deep learning, namely text classification based on aspect sentiment analysis, baseline deep learning methods, capsule networks and attention mechanism. Details about the methodology used, proposed system and evaluation measures are mentioned in the fourth section, and in the fifth section all results and analysis of the research performed are included. Finally, we give a conclusion in the sixth section.

2 Definitions

2.1 Convolutional Neural Networks

A ConvNet, or CNN is categorized in the form of a deep, feed-forward artificial network, or ANN which is inculcated majorly to refine the processing to recognise image to voice recognition [10], or in classification, segmentation. In this model, it encompasses neurons, represented by three dimensions, the spatial dimensionality of the input (height and the width) and the depth, which are self-optimizing in nature. It is a type of ANN that usually comprises a set of layers namely, convolutional, fully connected, and pooling layers. Each of these layers has unique parameters that can be used for optimization and will perform different tasks on input data.

1. Input layer - It will hold the input data of the dataset.
2. Convolutional layers - Convolutional layers, related to feature extraction are the type of layers where filters are applied and it comprises parameters such as a number of ‘kernels’ or filters and size of ‘kernels’ or filters. It is a type of layer where all the user-defined parameters are present and consist of filters called ‘kernels.
3. Pooling layer- It consists of performing the process of extracting a particular value from a set of values, specific operations of pooling are performed such as max, average, or min pooling to basically reduce the dimensionality of the network.

4. Fully connected layer- This layer forms the last block of the CNN architecture, related to the task of classification. It is a type of layer that takes input from the previous layer for the classification output of a CNN and is used to provide more flattened results.

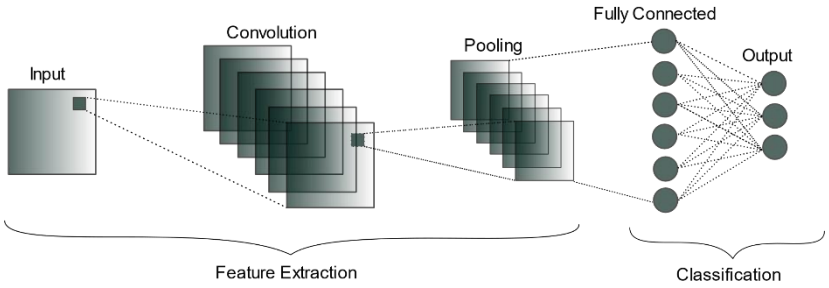

Figure 1 Framework for convolutional neural network

2.2 Recurrent Neural Networks

RNNs are a type of semantic network that has shown considerable potential in MT tasks [13] and is also a type of ANN that is majorly used in recognition techniques like handwriting [14] and speech [15]. Unlike feedforward semantic nets, these types of ANN have inputs connected to each other which will result in materialistic technology to enhance the behavior in the sequence of time that results in agreeable contrast for sentiment analysis. These RNN are a network with nodes in the form of neurons and follow a particular sequence that creates a directed graph in which the following pattern is observed by giving input using the backward nodes and resulting the output inside forward nodes. Also, these nodes have a real valued weight attached to them which will result in making modifications to the output and the signals passing through them. Several hidden states can be obtained by referring to an example sentence S, i.e. $H = [h_1, h_2, \ldots, h_n]$ by feeding the input $X = [x_1, x_2, x_3 \ldots, x_n]$ through RNN with dim_w amd dim_h, measure of word embedding and the latter denotes the hidden states.

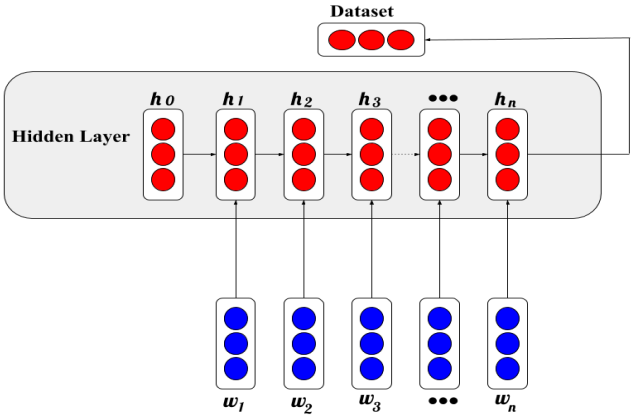

Fig. 2. Framework for recurrent neural network

Although RNN's have several shortcomings which will make them difficult to model and train such as vanishing gradient problem or exploding gradient problem or it becomes difficult to process very long sequences if we are using 'relu' or 'softmax' function as an activation function. the storage unit that is controlled by various networks, at the time of replacement might create errors like integrated time delays and feedback loops. Thus, we require a gated mechanism for RNN.

Look at figures 3(a) and 3(b) for illustrations of the gated RNN such as LSTM unit and GRU units respectively.

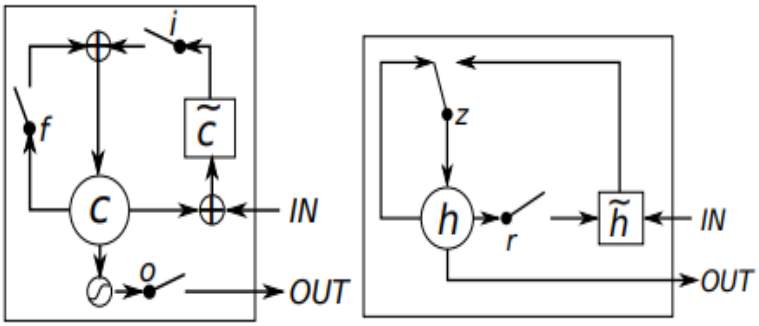

Fig. 3. (a) LSTM Unit [19] **(b)** GRU [19]

2.3 Long Short-Term Memory (LSTM)

Hochreiter and Schmidhuber contributed in the field of machine learning by developing the LSTM [5] unit. The memory cell in LSTM is used for storing the information. After several years of progress in the concerned field, addition of a forget gate along with many other changes were observed. Input, forget and output gate is used for management of memory. The main benefit of using a LSTM unit is that it remembers the flow of important features throughout the long distance to be covered. Despite its complex implementation, still it is used in various areas of machine learning that involves MT, analysis and pattern recognition.

For each function which is nonlinear in nature, every j^{th} LSTM unit preserves a memory c_t^i. The output h_t^j, thus the resulting activation function is:-

$$h_t^j = o_t^j \ tanh\ tanh\ (c_t^j) \tag{1}$$

where o_t^j is the output gate. Finally output gate is computed as,

$$o_t^j = \sigma(W_0 x_t + U_0 h_{t-1} + V_0 c_t)^j \tag{2}$$

where V_0 is a diagonal pattern and σ is a logistic sigmoid operation.
The memory cell c_t^j is modulated by deleting the previous content and followed by addition of $\underline{c_t^j}$,

$$c_t^j = f_t^j \ c_{t-1}^j + i_t^j \ \underline{c_t^j} \tag{3}$$

where the new memory content is defined as

$$\underline{c_t^j} = tanh\ (W_c x_t + U_c h_{t-1})^j \tag{4}$$

Forget gate f_t^j formulates the degree to which the main memory cell of LSTM is augmented by the new memory content and finally evaluated by an input gate i_t^j . Gates are finally evaluated by,

$$f_t^j = \sigma(W_f x_t + U_f h_{t-1} + V_f c_t)^j \tag{5}$$

$$i_t^j = \sigma(W_i x_t + U_i h_{t-1} + V_i c_t)^j \quad (6)$$

Note that V_f and V_i are diagonal matrices.

2.4 Gated Recurrent Unit (GRU)

Gated Recurrent Unit, or GRU, introduced by Cho, et al [6] is an extended version of LSTM networks. GRU only has an update gate, which is a combination of two gates i.e., forget and input gate, and a reset gate. This model is comparatively more straightforward and easier in computation as compared to standard LSTM models. GRU's got rid of the cell state that means it doesn't have any additional memory cells and thus uses the hidden state to transfer information.

Considering GRU at time t, the linear hindrance among the previous activation h_{t-1}^j and the new upcoming activation [11] h_t^j, is denoted as the activation h_t^j.

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \underline{h}_t^j \quad (7)$$

z_{ret}^j is denoted by an update gate which is responsible to measure the evaluation of extent of activation. Evaluation of update gate is,

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \quad (8)$$

Similar calculation in accordance to previous recurrent unit method is done for $\underline{h}_t^j$:

$$\underline{h}_t^j = tanh\,(W x_t + U(r_t \odot h_{t-1}) \quad (9)$$

where $\odot$ represents element-wise multiplication and r_t is defined as a calculated set of reset gates. The reset gate r_t^j is calculated in the same manner to the update gate,

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \quad (10)$$

2.5 Bidirectional-Gated Recurrent Unit (Bi-GRU)

To professionalize and improve the current state of training complexity, introduction of Bi-GRU [11] was involved. as it provides the best functionality to improve the computational cost of training when applied on larger datasets. It functions in both the directions of forward manner as well as reverse manner to make further predictions of the previously calculated state by applying the combination of cell state and hidden state. BiGRU belongs to the class of RNN which are bidirectional in nature, or a deep learning architecture with only forget and input gates.

2.6 Bidirectional-Long Short-Term Memory (Bi-LSTM)

As an extension to the conventional LSTM layer, BiLSTM layer [12] is an RNN, generative form of deep learning that came into existence in 1997 by Schuster and Paliwal which in turn make use of more than one LSTM which helps in learning expression of the sequence, one time in each direction of left and right and vice versa.. It is widely useful in applications like sentiment classification, sentence classification or handwriting recognition where the learning problem is sequential and there is a need to understand the context more efficiently. In order to get the final evaluated result of the output sequence, Bi-LSTM requires a set of more than one parallel layer that includes both backward and forward hidden layers. By taking advantage of this, the ultimate scenario will result in fewer delays as compared to unidirectional RNN. BiLSTM architecture comes into existence from their stamina to store long distance functioning inputs thus solve vanishing gradient problems.

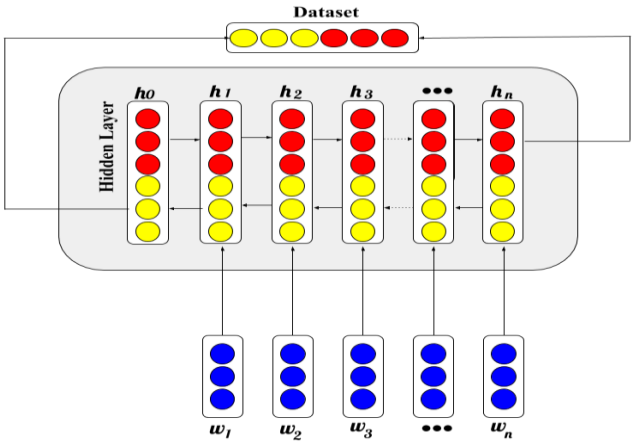

Fig. 4. Framework for Bidirectional Recurrent Neural Network (Bi-RNN)

2.7 Attention Mechanism

This type of mechanism (attention) was popularised by [16] which is widely used for improving various types of semantic nets or neural nets such as convNets, RNN's. Self-attention is the process in which we apply attention mechanisms in each concerned point of the main input sequence. In this type of attention, for each sequence position, we create 3 vectors namely, query, key, value. Thereafter we modulate and apply the outcome attention to each position x_i, using the x_i query vector, and key and value vectors for all other positions. Finally, the applied sequence X of words is transformed into a sequence Y where each y_i contains all the information about x_i .
The above three vectors can be created by applying linear projections which are learned in nature [17] or by making use of the type of layers which are feed forward. The final resulting order can be evaluated by grouping the parameters in Q, K, V matrices [17].

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{11}$$

2.8 Capsule Network

Capsule network, a type of ANN that first came into existence in 2011 by Geoffrey Hinton [18] as an alternative to CNN. The problem of max pooling and other DNN is solved by CapsNet as they give better results when talking about levelled hierarchical relationships. It is a type of network that contains numerous capsules which contain groups of neurons which are used to learn or detect an object. These neurons enclose all the relevant or crucial knowledge that is associated with the features such as position, size and hue and are represented in the vector form as an output. A squashing function is used to differentiate among several capsules and their designing networks and is applied along with the activation function.

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{12}$$

where v_j is the vector output of capsule j and s_j is its total input.

We define capsule layers by modelling in different levels out of which the last level denotes the primary capsule. These small regions try to predict shapes like any quadrilateral. The upper layers are signified by routing capsules. Capsules interact with each other through an iterative "routing-by-agreement" mechanism which was introduced by Geoffrey Hinton [7]. Also, capsNets have a unique feature in which lower-level capsules interact with higher-level capsules by sending them an output. Capsules are very promising in applications such as emotion detection, hand-written and text recognition, MT and many more.

3 Related Works

3.1 Text Classification based on aspect-based analysis

To justify the classification of polar sentences, a methodology was implemented based on the aspect-based analysis. Zhuang Chen et al. [21] worked on a model named TransCap which is basically used to conduct aspect level sentiment classification with document-level knowledge. Also, in this model, they used both the Aspect routing approach for encapsulating the semantic representations and Dynamic Routing approach to adaptively couple the semantic capsules. On the basis of the achieved results that was performed on the two major SamEval datasets, it was shown that the model successfully enhanced the previous accuracy of other baseline models.

Yunlong Liang et al. [22] prepared aspect-based model that works to improve the polar cases of text classification. For this, they worked on Aspect-guided-deep-Transition (AGDT), which consists of aspect-concatenated embedding, aspect-reconstruction, and aspect guided encoder. Further, they used several baseline models like CNN, TD-LSTM, IAN, RAM, and GCAE. The best performing outcome was from the DS model which resulted in 87.72%.

Tao Yang et al. [23] created a new target subnetwork to capture the main fundamental information which consists of conceptual and textual information. In regard to the previous studies where only a polar type of cues was used, these contributed by creating approximately five different types of lexical cues for utilizing the exact information accumulated through words. To utilize the dependency between several cues and words, the authors created a new attention mechanism. Further, they used various baseline models like LSTM, RAM, CueNet, etc. out of which the maximum accuracy obtained was 88.6%.

Yanyan Wang et al. [24] worked on enhancing the previous deep learning technologies and applied a new methodology of SenHint on the datasets of English and Chinese benchmark. These datasets were polar in nature due to which they performed sentiment-based analysis and implemented Markov-Logic Network. Finally, the outcome comparison was made in which SenHint-rel was found better than GCAE.

3.2 Analysis of datasets using baseline deep learning models

Various deep learning models are being used for classification purposes. These models in combination with several other baseline models serve features and improvised vector functions to enhance the accuracy and reduce dimensions. These models are also used in mapping sequences of outputs.

Ashraf Elnagar et al. [25] proposed a new single label corpus called the SANAD corpus which contains 200k articles so as to perform text categorization on Arabic news articles. Along with the available categorization systems, several deep learning classifiers were also introduced so as to reduce the need

for preprocessing. In this paper, they implemented 9 models namely CNN, RNN, and attention-based models. The results draw a conclusion that DNN models performed best on The SANAD corpus (maximum with HANGRU 95.18% and minimum with CGRU 93.43%).

Tanjim Ul Haque et al. [26] used a supervised learning method on the amazon product reviews dataset to polarize those reviews and get better results. In this model, they use both approaches - active and manual for the labeling of datasets. The datasets were divided on the basis of three categories- Accessories and Musical Instruments, Electronic and Cell Phone Reviews. To perform feature extraction, they used Chi-square and tf-IDF in order to enhance the accuracy. The results show that 10-fold provided better accuracy than 5-fold and also a linear support vector machine provides the best results with 94.02% accuracy.

Anirban Mukherjee et al. [27] used different types of DNN's on the dataset to figure out the best model that outperforms the remaining models, as they were targeting sentiment analysis using the concept of multi class classification. Further, they applied oversampling on the dataset to show the improvement inaccuracy. In this experiment, they observe that bidirectional LSTM shows the best results without oversampling and after oversampling, the results show a significant increase with 80% accuracy.

Ryan Ong [28] started experimenting by testing out different variations of RNN models including LSTM and GRU both unidirectional and bidirectional with CNN layers. Also, he used different kinds of word embeddings to see if there is a significant change in the results. The results show that BiLSTM-CNN shows significant results and therefore they conducted a manual search for some of the key hyperparameters like the level of epochs, spatial dropout rate and Pre-trained vs No pre-trained embeddings. Also, from this experiment conclusion was made that ordering of layers is very important and shows a significant change in the results.

Sharat Sachin et al. [29] conducted a survey for different DNN on the amazon reviews dataset. In this model, they have 3 layers- embedding, unidirectional/bidirectional and dense layers. In this paper, the conclusion was drawn that bidirectional performs better than unidirectional and also GRU's are faster to train than LSTM. BiGRU outperforms all other RNN techniques in terms of accuracy.

Gonzalo A. Ruz et al. [30] proposed a methodology on twitter dataset with a motive to analyse it by using a network classifier called Bayesian. In this dataset, they classified the sentences into positive type tweets and negative type tweets. Along with the BF-TAN, they deployed a model of SVM with random forest classifiers to compare the performance on Spanish texts as well. To tackle the problem of class imbalance they used SMOTE features. The final result was enhanced by the performance of SVM that increased their accuracy above 80% compared to all the other modules present in the processing part.

3.3 On the basis of Attention mechanism

While implementing sequence modelling networks, the entire details and information regarding input is contained inside the current state of various CNN and LSTM, BIGRU. There is no doubt that this information is precise and complete. Assumption mechanism deploys the ease of looking forward in

the entire hidden state of input and helps in enhancing the computational accuracy. Various researchers have worked upon utilizing the attention mechanism.

Liang Zhou et al. [31] presented a new model BiGRU attention-based model to perform text classification on Chinese texts to solve the problem of insufficient text context information learning and weak feature extraction ability. Their proposed model was divided into three parts: the input layer, hidden layer, and output layer. The input layer consisted of the CBOW model in word2vec. The hidden layer contains BIGRU. The output layer shows the output of BIGRU calculated by the softmax function. The accuracy attained by this model was 90.45% and it outperforms the other existing models.

Similarly, Jingren Zhang et al. [32] restructured the formulation of CNN models that are generally used as a deep learning module for text classification. They applied their new proposed mechanism on MRD and SamEval2016 datasets. The new fusion model consists of MATT(CNN) along with the combination of Bi-directional GRU (Bi-GRU). For dimensionality reduction they applied Principle Component Analysis (PCA) This improved their accuracy by 5.94% on the MRD dataset and 11.01% SamEval2016 dataset. The highest accuracy was 79.22% in the laptop category.

Yang Liua et al. [33] introduced an attention gated layer to amplify the output produced by CNN models for textual extraction. According to the analysis and results section, their model enhanced the accuracy of other standard CNN models by 3.1%. Along with this moving forward in their experiment, they applied several activation functions to outperform the existing Relu, sigmoid, softplus, PReLu, LReLu, etc. the proposed functionality that they formulated was Natural Logarithm rescaled Rectified Linear Unit (NLReLU). The final outcome result was 94.5% in the TREC dataset.

Weijiang Li et al. [34] worked in improving the functioning of previous deep learning baseline models for sentiment analysis. This field includes aspect-based terms also and can easily manipulate the outcome of any model. Their methodology provided an improvised version of bidirectional baseline models along with multi-channel features accommodation. They used MF-BiLSTM and SA-BiLSTM on MR, SST-5, SST-2 datasets. They achieved the highest accuracy of 89%.

Artaches Ambartsoumian et al. [35] worked on a new model based on self-attention networks for sentiment analysis called the SSAN model. They tested the proposed model on 2 layers of stacked configurations. The final model that produces capturing of long-term dependencies compared to the other model was SSAN. It achieves a better accuracy against CNN or RNN models also.

Likun Ji et al. [36] proposed a new model for sentiment analysis, which was a combination of self-attention mechanism with BiGRU. In this paper, they introduced multi head self-attention along with two embeddings used i.e., position and word. They evaluate the proposed model on IMDB dataset which shows 90% accuracy and thus outperforms other models and also the results show that this model is more effective in extracting information for sentiment analysis.

3.4 Capsule Networks for text classification

Various types of important and valuable information is lost while performing the max-pooling stage in CNN in the process of text categorization. Therefore, keeping this in mind, a new technology of capsule networks was introduced. Capsule networks were basically made to target the lower level features using routing by agreement strategy. Many researchers have compared their several experiments and implemented capsule networks in their datasets, showing the enhanced accuracy based on the results achieved.

Haftu Wedajo Fentaw et al. [37] along with his team worked on the CapsNets Model. Comparison and analysis were done of the resultant model with several DNN models along with baseline models on the five real-world datasets which show significant results. In this paper, they made many observations which were used to enhance the computational complexity, also CapsNets models are quite sensitive regarding the sequence in which the words are arranged which means CapsNet models can work better while capturing spatial information.

Dynamic routing enhancement was the most important key factor that catalyzes the entire functioning of a capsule network. Wei Zhao et al. [38] explored CapsNet for text classifications by using dynamic routing. In the proposed model they used 4 layers. They conduct the experiment on 6 benchmarks which includes various text classification tasks. Also, they transferred a single label to multi-label text classifications which shows significant improvements in the results over the baseline methods.

Hongxia Yin et al. [39] worked on Capsule network model in notion to perform classification via cross-domain along with Identifying Transferable Knowledge (CITK). They also use BERT for pre-training. They performed the proposed model on a dataset that consisted of 2 domains: electronic and kitchen domain- positive and negative. According to the output, the CITK model outperforms other methods significantly.

Along with the approach to use capsule networks, various researchers implemented these networks along with the aspect-based mechanisms and deep learning strategies. Chi Xu et al. [40] developed a model named CAPSAR (a capsule net model) for improvisation of the aspect level analysis. The capsule was used to denote the sentiment categories and then with the help of sentiment-aspect reconstruction procedure they inject aspect type information into sentiment capsules. The model is performed on 3 real-world benchmarks and shows superiority over the other models. To enhance the accuracy by using deep learning layers, it helps in accumulating the important features that are compared with training labeled datasets.

Yongping Du et al. [41] created a hybrid NN based on CapsNet with addition to BiGRU that is used to obtain the implicit sentiment information and achieve features over long distances. The model is performed on two text datasets and obtained an accuracy of 82.5%. Also, in this model, they used a self-attention mechanism along with CNN to reduce timing and implementation cost.

Kejun Zhang et al. [42] along with this team proposed a SC-BiCapsNet for classification and analysis. In the proposed model there are basically 3 parts of modules, G-word vector and Bi-channel capsule representation. This paper used 2 datasets of IMDB and NLPCC 2014 and the results show that the developed method gave better results compared to other methods on the IMDB dataset with an accuracy of 92.38%.

Jaeyoung Kim et al. [43] compared the performance between traditional DNN along with the capsules used for text classification. They used several datasets such as news, MR’s, etc., Further, they represented a simple routing method to outperform the computational complexity created by effective routing mechanisms which are dynamic in nature in the capsNets. The final result achieved was 87% as trained by simple routing.

Xian Zhong et al. [44] worked in analyzing the difference between the traditional deep learning ANN methods with the latest capsule network technology. Based on the incorrect compression mechanisms, the team proposed an IPC-CapsNet algorithm which was combined with SPT-CapsNet so as to improve the results in terms of the complexity. The final complexity along with accuracy were improved to 85.89%, which was 84.90% earlier.

Deepak Kumar Jain et al. [45] suggested a methodology in order to check the sincerity of all those sites and communities that provides questions and answers. This project was made to remove irrelevant content and improve the quality of the concerned material. They applied a deep refinement methodology which comprises C-BiLSTM. The resultant model outperformed the capacity and accuracy of SVM and Bi-LSTM. The final result produced the highest accuracy of 96.03%.

Min Yang et al. [46] investigated the converting capability in the field of text classification. Their capsNet model outperformed the performance of different models on 4 out of 6 datasets. They used capsule compression and class-guided routing in the dynamic routing mechanism of the final capsule model. They achieved the highest accuracy of 82.3%. Further, they investigated various combinations of baseline models along with capsule networks.

4 Methodology

4.1 Dataset Used

The dataset used is a large set of reviews from Amazon developed with work from Jianmo Ni [47], which contains the content of reviews, along with some metadata that is labelled with a 5-point system, 1 being the worst and 5 being the best. A 5-core subset of the dataset has been used, which means that each of the users and items present in the subset have at least 5 reviews. This ensures a higher quality of reviews. We have carried out our analysis on reviews from the Electronics section, extracting only 120 thousand reviews which is a reasonable amount. For purposes of getting a balanced dataset, we use 40 thousand reviews from each category, namely ‘negative’, ‘neutral’ and ‘positive’. This is what we use for the 3-class part of the experimentation. We have characterized the negative review part as 1 or 2, the neutral one is number 3 and positive review falls under the category of 4,5. Finally, splitting this into 90 thousand reviews for training and 30 thousand reviews were used for testing.
For the 2-class variant, we have only two classes- positive and negative. We choose 100 thousand reviews from each class (200 thousand reviews total) and split this into 2 parts - 150 thousand for training and 50 thousand for testing.

4.2 Proposed System

In this study, we have listed works that focus on sentiment analysis. We further wish to compare models prepared by adding and removing layers and mechanisms that have been shown in the listed papers to improve accuracy of the classifiers that perform sentiment analysis. Thus, we try to add a CNN layer and an Attention layer. We also demonstrate the use of capsule mechanisms to classify reviews. We perform this on LSTM, Bi-LSTM, GRU and Bi-GRU baseline models.

For the 3-class model, we selected 40 thousand reviews from the 'negative' class (1-2 stars), 'neutral' class (3 stars) and the 'positive' class (4-5 stars). The total number of reviews was thus 120 thousand. Regarding the 2-class model, we selected 100 thousand reviews from both the 'negative' and 'positive' class, the total being 200 thousand. Now we performed preprocessing on these reviews, with steps like tokenization, removing punctuation. We retain and use the 20 thousand most common, thus not bringing the rare words into account for sentiment analysis. Then the content of the review is mapped to a list of

integers, limiting the number of integers in the sequence to 150. This is because by limiting review length to 150, we get the full text of 90% of the reviews. The rest of the sequence is padded with a neutral integer. For the purpose of preparing the target variables of the data, we just convert them into one-hot encodings for the 3-class model and keep them as a single binary variable for the 2-class model.

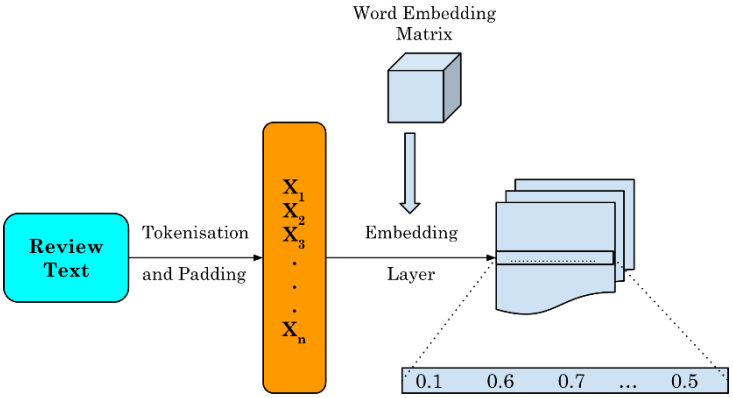

Fig. 5. Preprocessing steps for reviews

The implementation of the models was done in Keras, with word vector embedding using the GloVe pre-processed weights being performed before they were fed into the models. To prepare the matrix of word embeddings, we get the pretrained weights from the GloVe data file, and we replicate the vectors of words we have used to fill up the embedding matrix. We use the embeddings of dimension 300, implying that each single word in the review is represented by a vector of length 300. This is passed as one of the parameters to the model we create.

The first layer of all the models is the Embedding layer, to which one of the parameters passed is the embedding matrix we created previously. Each review is converted into its matrix form by this layer.

The various models are explained in figure 6. For the models involving the attention and CNN layers, we had the following layers: the Conv1D is the first layer with 128 filters, a kernel size of 5 and 'relu' activation, preceding the GRU / LSTM layer (unidirectional or bidirectional), with a dropout regularization of 0.1 being considered for the layer. After considering and experimenting with several other output sizes, 128 was the number which was analysed as the exact fit in the entire structure. Next, we have a SeqSelfAttention layer that we use with 'sigmoid' activation, and finally we have a Flatten layer to flatten the output. The last layer that acts as an output layer is the Dense layer with 3 outputs enhanced by categorical cross entropy loss and the softmax activation function.

For the capsule model, the unidirectional or bidirectional layer is followed by the capsule layer block, with 10 capsules, the dimensions being 16 and the number of routings being 5. Then we flatten the output of this layer, finally having a Dense layer with a single output with 'sigmoid' activation.

The mechanism of early stopping was also used to stop the training after the performance of the model stops improving on testing data.

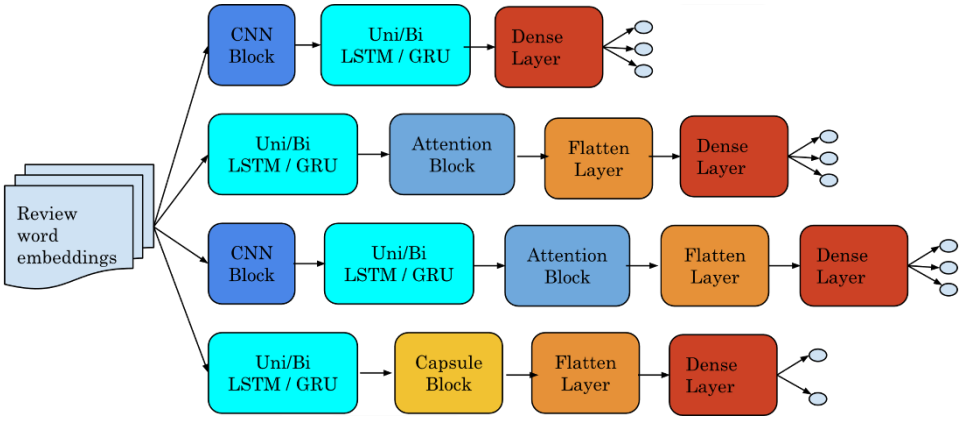

Fig. 6. Layers for the various models

4.3 Evaluation Measures

These types of evaluation metrics are important in order to measure the performance of classification. We have various sampling criteria on the basis of which we can analyse these metrics:

1. True positives denoted as TP shows sum of data accurately classified.
2. False positives denoted as FP shows sum of data accurately misclassified or predicted.
3. False negatives denoted as FN shows sum of wrongly classified data as correct.
4. True negatives denoted as TN shows the sum of misinterpreted data predicted.

Other parameters involved are:

4.3.1 Accuracy

It is the most common type of metric used for comparing and evaluating the performance of classification. It represents the proportion of the accurate predictions; evaluation can be done as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

Though, it can be a useful measure of the performance of a classifier in maximum situations, it is not enough to give a proper decision. Therefore, other measures are considered for better evaluations.

4.3.2 Precision

Precision is characterized as the ratio of all the data that is correctly predicted and successfully compared to the data that is incorrectly classified.

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

For the variables and features that are in sparse manner, they can be easily classified based on precision.

4.3.3 Recall

Recall is determined as the fraction of the actual data which is correctly classified among the samples compared to the one that are correctly classified along with those who were not in a particular category

but still, they were not classified. It shows the actual cases of those samples which are not in the required category of the outcome.

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

4.3.4 F1-Score

F1-Score is defined as the perfect criteria for classification as it is a weighted harmonic mean combination of recall and precision.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

5 Result and analysis

5.1 Abbreviations used

Rec. – Recall
Acc. – Accuracy
F-sc. – F-score
Att. – Attention
Prec. – Precision
ANN – Artificial neural network
MT – Machine translation
DNN – Deep Neural Network
CapsNet – Capsule Neural Network
Bi-GRU – Bidirectional Gated Recurrent Unit
GRU – Gated Recurrent Unit
LSTM – Long Short-Term Memory
Bi-LSTM – Bidirectional Long Short-Term Memory

5.2 Effects of CNN and attention layers

We have tested the effects of convolutional, attention and combinations of them on unidirectional and bidirectional LSTM and GRU networks. There are 4 models for each type of network. The first one is the base model. The second contains the base model in conjunction with CNN layer. The third model contains the base model in combination with a self-attention layer. The fourth variant of the network uses both convolutional and self-attention layers in combination with the base model.

In table 1, we have compared the different models for LSTM, i.e., the base model, and the base model augmented with convolutional neural networks, attention layers and both of them in tandem. The attention augmented model works best for the 3-class model whereas the base model works best for the 2-class model.

Table 1. LSTM

	3 Class				**2 Class**			
	Acc.	**Prec.**	**Rec.**	**F-sc.**	**Acc.**	**Prec.**	**Rec.**	**F-sc.**
Base	70.05	70.30	70.30	70.30	**93.74**	**94.00**	**94.00**	**94.00**
Base+CNN	73.58	73.33	74.00	73.66	93.47	92.50	92.50	92.00
Base+Att.	**74.91**	**75.00**	**75.00**	**75.00**	92.57	93.50	93.50	93.50
Base+C+A	72.91	73.00	73.00	73.33	92.78	92.50	92.50	93.00

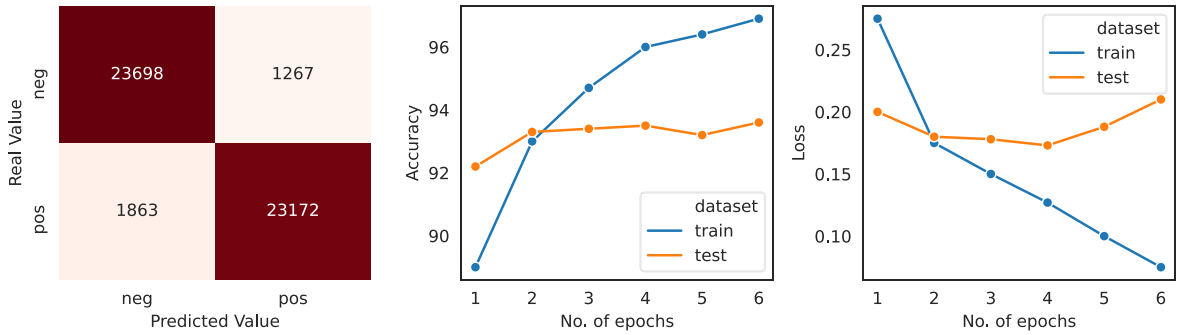

Fig. 7. Confusion matrix, accuracy, loss graphs for 2 class LSTM

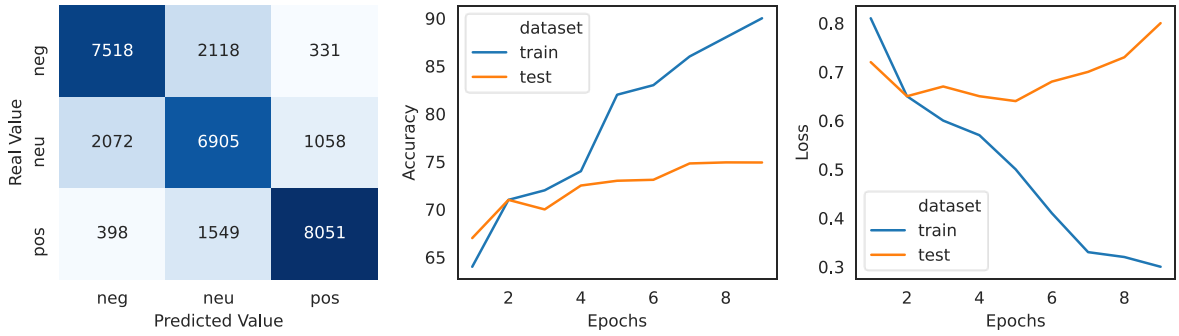

Fig. 8. Confusion matrix, accuracy, loss graphs for 3 class LSTM with attention

In table 2, we have compared the different models for GRU, i.e., the base GRU model, and the GRU model augmented with convolutional neural networks, attention layers and both of them in tandem. The CNN augmented model works best for the 3-class model whereas the base model works best for 2-class model.

Table 2. GRU

	3 Class				**2 Class**			
	Acc.	**Prec.**	**Rec.**	**F-sc.**	**Acc.**	**Prec.**	**Rec.**	**F-sc.**
Base	71.06	71.30	71.00	71.30	**92.57**	**93.00**	**93.00**	**93.00**
Base+CNN	**72.42**	**73.33**	**73.00**	**73.00**	92.17	92.50	92.00	92.00
Base+Att.	71.23	71.33	71.33	70.33	92.48	92.50	92.50	92.50
Base+C+A	72.09	73.00	73.33	73.00	92.51	92.50	92.50	92.50

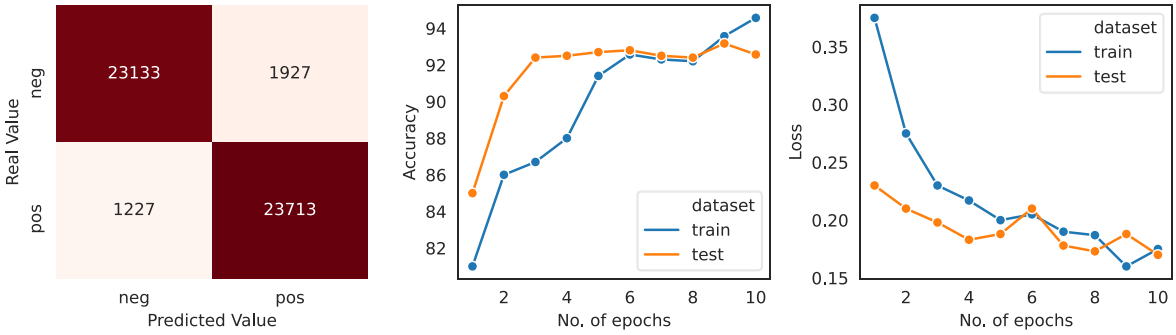

Fig. 9. Confusion matrix, accuracy, loss graphs for 2 class GRU

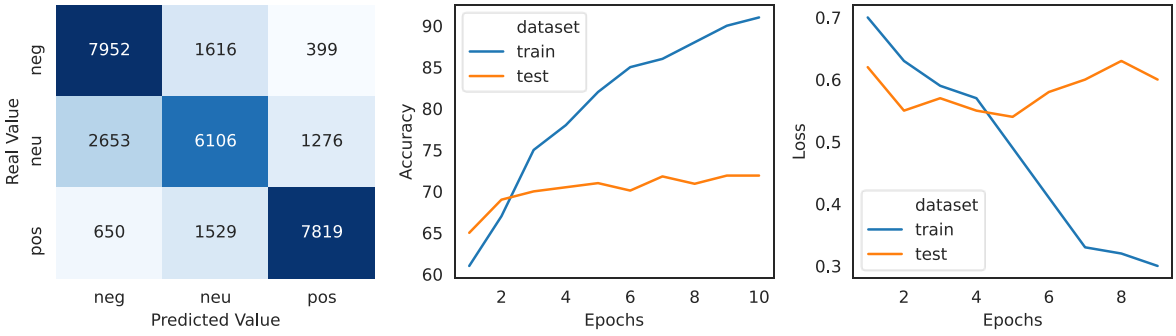

Fig. 10. Confusion matrix, accuracy, loss graphs for 3 class GRU with CNN

In table 3, we have compared the different models for Bi-LSTM, i.e., the augmentation of Bi-LSTM and the concerned base model along with convolutional neural networks, attention layers and both of them working together. The attention augmented model works best for the 3-class model whereas the base model works best for the 2-class model.

Table 3. Bi-LSTM

	3 Class				**2 Class**			
	Acc.	**Prec.**	**Rec.**	**F-sc.**	**Acc.**	**Prec.**	**Rec.**	**F-sc.**
Base	70.42	70.30	70.30	70.3	**93.68**	**93.50**	**93.50**	**94.00**
Base+CNN	73.46	73.66	73.33	73.33	92.18	92.50	92.00	92.00
Base+Att.	**74.96**	**74.33**	**75.00**	**75.00**	93.56	94.00	93.50	93.50
Base+C+A	71.75	72.66	72.66	73.00	92.56	92.50	93.00	92.50

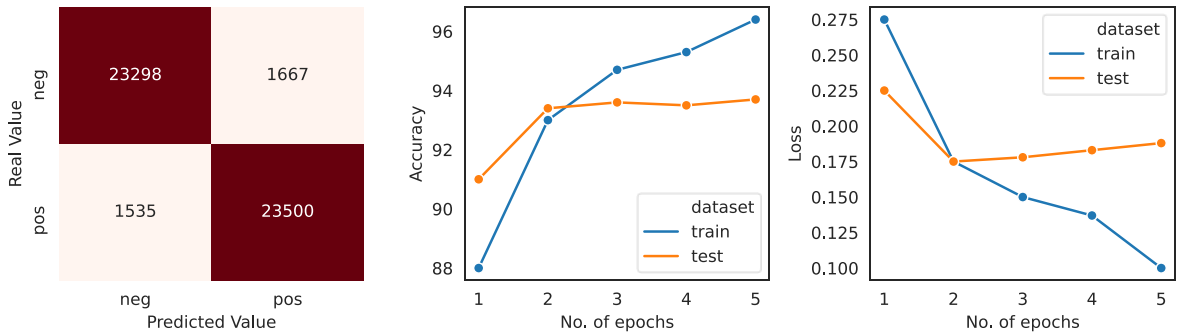

Fig. 11. Confusion matrix, accuracy, loss graphs for 2 class Bi-LSTM

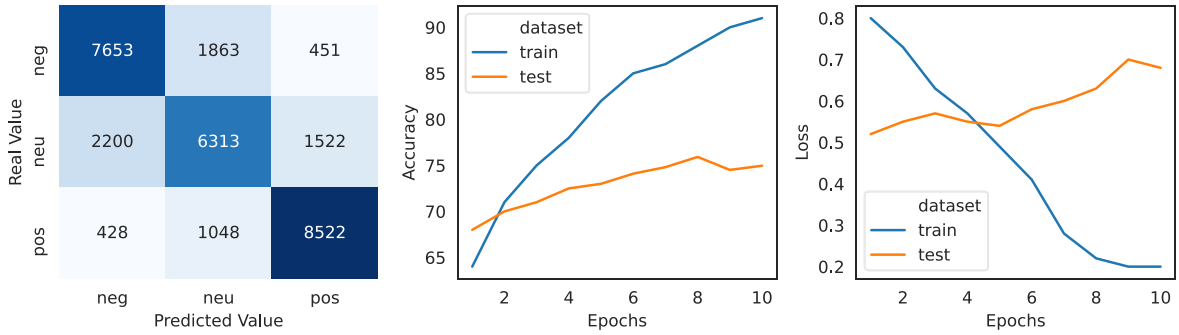

Fig. 12. Confusion matrix, accuracy, loss graphs for 3 class Bi-LSTM with attention

In table 4, we have compared the different models for Bi-GRU, i.e., the base model, and the Bi-GRU model augmented with convolutional neural networks, attention layers and both of them working together. The CNN augmented model works best for the 3-class model whereas the base model works best for 2-class model.

Table 4. Bi-GRU

	3 Class				**2 Class**			
	Acc.	**Prec.**	**Rec.**	**F-sc.**	**Acc.**	**Prec.**	**Rec.**	**F-sc.**
Base	71.19	71.00	70.06	70.06	**93.25**	**93.00**	**93.00**	**93.00**
Base+CNN	**72.86**	**73.00**	**73.00**	**73.00**	93.00	93.00	93.00	93.00
Base+Att.	71.00	70.00	70.00	70.00	91.62	92.00	91.50	91.50
Base+C+A	71.00	73.33	70.3	70.2	92.08	92.00	92.00	92.00

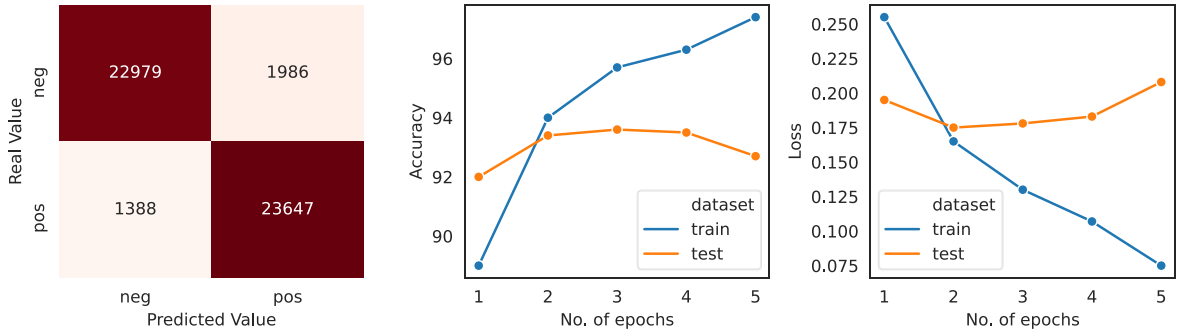

Fig. 13. Confusion matrix, accuracy, loss graphs for 2 class Bi-GRU

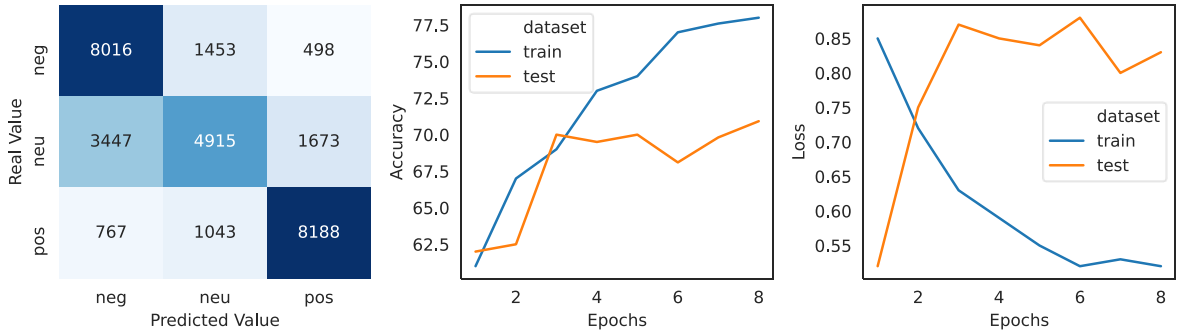

Fig. 14. Confusion matrix, accuracy, loss graphs for 3 class Bi-GRU with attention

For LSTM, Bi-LSTM and Bi-GRU networks, they appear to work best when augmented by an attention layer. The attention layer is one component of a network architecture that is in charge of managing and quantifying the interdependence between the input layers. This assigns priority to various values that signify sentiment in a review text and help improve the accuracy of these classifiers.

For GRU network, it is best augmented by a CNN layer before it. This layer convolves the input into an intermediate representation, which contains dimensions each of which can be considered an extracted feature. The GRU then takes these as input. The GRU is a simpler structure with lesser parameters to train, in theory this allows it to generalize better on data, and allows it to better work with the output of the CNN layer.

The 2 class models for LSTM, Bi-LSTM, GRU and Bi-GRU do not seem to improve their accuracy with augmentation of convolutional and self-attention layers. This can be assumed to be because the higher number of layers increases the complexity of the model, which may not have the best setting for optimal results, the simplicity of the base models thus having the highest accuracy.

5.3 Effect of Capsule Network

Table 5. Metrics for different models with and without capsule block

		Acc.	**Prec.**	**Rec.**	**F-sc.**
LSTM	**Base**	93.74	94.00	94.00	94.00
	Base + Capsule	94.18	94.50	94.50	94.50
GRU	**Base**	92.57	93.00	93.00	93.00
	Base + Capsule	93.83	93.50	93.50	94.00
Bi-LSTM	**Base**	93.68	93.50	93.50	94.00
	Base + Capsule	**94.87**	**95.00**	**95.00**	**95.00**
Bi-GRU	**Base**	93.25	93.00	93.00	93.00
	Base + Capsule	93.77	94.00	94.00	94.00

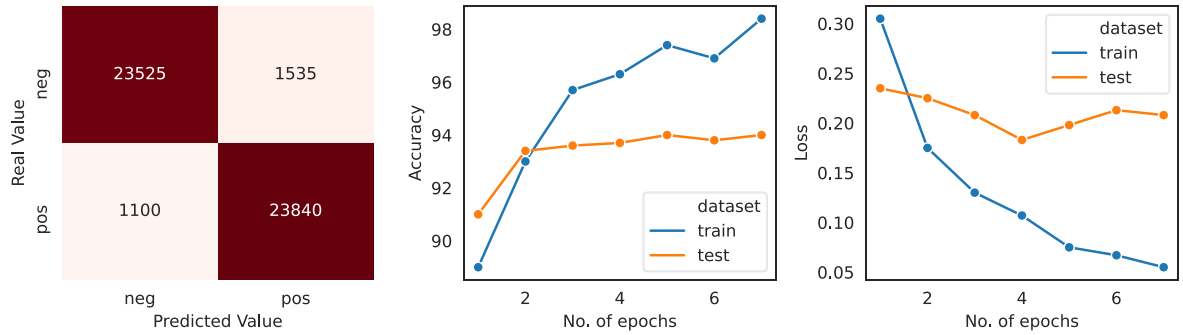

Fig. 15. Confusion matrix, accuracy, loss graphs for 2 class Bi-LSTM with Capsule

We finally also compare the capsule network by combining it with unidirectional and bidirectional LSTM and GRU, thus finding that it shows the best results with Bi-LSTM. The bidirectional nature of the Bi-LSTM provides a greater potential when fed with the experience or training of capsule block, thus increasing the ability for learning and generalization of the model in order to enhance the accuracy in the experiments performed. The Bi-LSTM is the most complex model containing the greatest number of parameters. It thus performs best when used in combination with the capsule block.

6 Conclusion

After analyzing all these results obtained by us, it is clear that capsule network and convolutional and attention layers have their inherent advantages when used in comparison with base models of unidirectional and bidirectional LSTM and GRU. Creating understanding from text is a subdiscipline of natural language processing that has many applications. The combination of CNN layers for spatial feature extraction along with LSTM / GRU cells to learn the temporal features is a classic example of the whole being better than the sum of its parts. Attention layers, which help to selectively focus on one aspect of the text while ignoring others can help with this goal as text has inherent structure that makes some parts more important than others. Capsule networks are useful for learning higher level features and details in an object. These combinations of models can be useful for other sentiment tasks. In the future, we might try to extract more information from these reviews, like specific emotions being felt by the reviewer. These might require further enhancements to the models. One other path would be to try these models on data from other places, like tweets or forum comments from different sites.

References

1. Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102-107.
2. Chaturvedi, S., Mishra, V., & Mishra, N. (2017, September). Sentiment analysis using machine learning for business intelligence. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 2162-2166). IEEE.
3. Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. Tourism Management, 61, 43-54.
4. Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015, June). An empirical exploration of recurrent network architectures. In International conference on machine learning (pp. 2342-2350).
5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
6. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)
7. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in neural information processing systems (pp. 3856-3866).
8. Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6), 424.
9. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.
10. Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). IEEE.
11. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480-1489).
12. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. CoRR Journal, arXiv:1508.01991.
13. Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2014. A recursive recurrent neural network for statistical machine translation. In ACL (1), pages 1491–1500.
14. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. IEEE transactions on pattern analysis and machine intelligence,31(5), 855-868.
15. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.
16. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In International conference on learning representations, 2015.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
18. Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011, June). Transforming auto-encoders. In International conference on artificial neural networks (pp. 44-51). Springer, Berlin, Heidelberg.
19. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Workshop on Deep Learning, December 2014.
20. Phung VH, Rhee EJ. A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. Applied Sciences. 2019; 9(21):4500.
21. Chen, Z., & Qian, T. (2019, July). Transfer capsule network for aspect level sentiment classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 547-556).
22. Liang, Y., Meng, F., Zhang, J., Xu, J., Chen, Y., & Zhou, J. (2019). A Novel Aspect-Guided Deep Transition Model for Aspect Based Sentiment Analysis. EMNLP/IJCNLP, 2019.
23. Yang, T., Yin, Q., Yang, L., & Wu, O. (2019). Aspect-based Sentiment Analysis with New Target Representation and Dependency Attention. IEEE Transactions on Affective Computing.
24. Wang, Y., Chen, Q., Ahmed, M., Li, Z., Pan, W., & Liu, H. (2019). Joint Inference for Aspect-level Sentiment Analysis by Deep Neural Networks and Linguistic Hints. IEEE Transactions on Knowledge and Data Engineering.

25. Elnagar, A., Einea, O., & Al-Debsi, R. (2019). Automatic text tagging of Arabic news articles using ensemble deep learning models. In Proceedings of the 3rd International Conference on Natural Language and Speech Processing (pp. 59-66).
26. Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE International Conference on Innovative Research and Development (ICIRD) (pp. 1-6). IEEE.
27. Mukherjee, A., Mukhopadhyay, S., Panigrahi, P. K., & Goswami, S. (2019, October). Utilization of Oversampling for multiclass sentiment analysis on Amazon Review Dataset. In 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (pp. 1-6). IEEE.
28. Ong, R. (2019). Offensive Language Analysis using Deep Learning Architecture. CoRR Journal, arXiv:1903.05280.
29. Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment Analysis Using Gated Recurrent Neural Networks. SN Computer Science, 1(2), 1-13.
30. Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92-104.
31. Zhou, L., & Bian, X. (2019, November). Improved text sentiment classification method based on BiGRU-Attention. In Journal of Physics: Conference Series (Vol. 1345, No. 3, p. 032097). IOP Publishing.
32. Zhang, J., Liu, F. A., Xu, W., & Yu, H. (2019). Feature Fusion Text Classification Model Combining CNN and BiGRU with Multi-Attention Mechanism. Future Internet, 11(11), 237.
33. Liu, Y., Ji, L., Huang, R., Ming, T., Gao, C., & Zhang, J. (2019). An attention-gated convolutional neural network for sentence classification. Intelligent Data Analysis, 23(5), 1091-1107.
34. Li, W., Qi, F., Tang, M., & Yu, Z. (2020). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. Neurocomputing.
35. Ambartsoumian, A., & Popowich, F. (2018). Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers. WASSA@EMNLP, 2019.
36. Ji, L., Gong, P., & Yao, Z. (2019, March). A text sentiment analysis model based on self-attention mechanism. In Proceedings of the 3rd International Conference on High Performance Compilation, Computing and Communications (pp. 33-37).
37. Fentaw, H. W., & Kim, T. H. (2019). Design and investigation of capsule networks for sentence classification. Applied Sciences, 9(11), 2200.
38. Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., & Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. (pp. 3110-3119)
39. Yin, H., Liu, P., Zhu, Z., Li, W., & Wang, Q. (2019). Capsule Network with Identifying Transferable Knowledge for Cross-Domain Sentiment Classification. IEEE Access, 7, 153171-153182.
40. Xu, C., Feng, H., Yu, G., Yang, M., Wang, X., & Ao, X. (2019). Discovering Protagonist of Sentiment with Aspect Reconstructed Capsule Network. CoRR Journal, arxiv:1912.10785.
41. Du, Y., Zhao, X., He, M., & Guo, W. (2019). A novel capsule-based hybrid neural network for sentiment classification. IEEE Access, 7, 39321-39328.
42. Zhang, K., Jiao, M., Chen, X., Wang, Z., Liu, B., & Liu, L. (2019). SC-BiCapsNet: A Sentiment Classification Model Based on Bi-Channel Capsule Network. IEEE Access, 7, 171801-171813.
43. Kim, J., Jang, S., Park, E., & Choi, S. (2020). Text classification using capsules. Neurocomputing, 376, 214-221.
44. Zhong, X., Liu, J., Li, L., Chen, S., Lu, W., Dong, Y., ... & Zhong, L. (2019). An emotion classification algorithm based on SPT-CapsNet. Neural Computing and Applications, 1-15.
45. Jain, D. K., Jain, R., Upadhyay, Y., Kathuria, A., & Lan, X. (2019). Deep Refinement: capsule network with attention mechanism-based system for text classification. Neural Computing and Applications, 1-18.
46. Yang, M., Zhao, W., Chen, L., Qu, Q., Zhao, Z., & Shen, Y. (2019). Investigating the transferring capability of capsule networks for text classification. Neural Networks, 118, 247-261.

47. Ni, J., Li, J., & McAuley, J. (2019, November). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188-197).