An Empirical Evaluation of various Word Embedding Models for Subjectivity Analysis Tasks

Ritika Nandi[1], **Geetha Maiya**[2], **Priya Kamath B**[3]
Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal Academy of Higher Education
Manipal, 576 104, India
[1]*ritika.nandi@learner.manipal.edu*,
{[2]*geetha.maiya*, [3]*priya.kamath*}*@manipal.edu*

Abstract

It is a clearly established fact that good categorization results are heavily dependent on representation techniques. Text representation is a necessity that must be fulfilled before working on any text analysis task since it creates a baseline which even advanced machine learning models fail to compensate. This paper aims to comprehensively analyze and quantitatively evaluate the various models to represent text in order to perform Subjectivity Analysis. We implement a diverse array of models on the Cornell Subjectivity Dataset. It is worth noting that the BERT Language Model gives much better results than any other model but is significantly computationally expensive than the other approaches. We obtained state-of-the-art results on the subjectivity task by fine-tuning the BERT Language Model. This can open up a lot of new avenues and potentially lead to a specialized model inspired by BERT dedicated to subjectivity analysis.

1 Introduction

The Internet has become increasingly accessible over the years with user numbers growing at an extremely fast rate. This has also led to a rapid rise in the number of people using various services online as well as registering their presence on various social networking platforms. The rapid spike in user numbers means that data is being generated at an unprecedented rate. Organizations are rapidly evolving their approach in order to utilize this data and are also trying to find sustainable ways to manage it. A major chunk of this data exists in the form of sequential textual data. Computer systems are well equipped to handle numerical data and perform well with numerical databases but this new form of data being generated necessitates the need for the development of specialized algorithms that convert this textual data into a form that can be understood by a machine (Bastas et al., 2019).

The most widely researched areas under Natural Language Processing (NLP) are tasks like Sentiment and Subjectivity Analysis, Machine Translation, and Automatic Question-Answering (Young et al., 2018; Sharma et al., 2020). One very interesting challenge posed due to the sheer volume of text generated these days is performing Subjectivity Analysis as a preliminary step before performing Sentiment Analysis, as filtering out statements that do not state an opinion or emotion reduces the time and resources required to perform Sentiment Analysis. We have taken upon this task, i.e., Subjectivity Analysis (Liu et al., 2010), and have used suitable metrics to evaluate the performance of each type of text representation method on the given dataset.

Subjectivity analysis recognizes the contextual polarity of opinions, attitudes, emotions, feelings etc. regarding products, services, topics, or issues. Subjectivity classification categorizes the given text as subjective or objective. While an objective text contains one or more facts about a product or an issue, a subjective text expresses the author's opinions (Karamibekr and Ghorbani, 2013).

2 Background

2.1 Previous Work

Subjectvity Analysis is a sub-task of Sentiment Analysis. Although, extensive research has been done in the field of Sentiment Analysis, limited study has been performed on Subjectivity Analysis. One of the earlier works in Subjectivity Analysis using the Cornell Subjectivity Dataset v1.0 (SUBJ) (Pang and Lee, 2004) is Self-Adaptive Hierarchical Sentence Model (AdaSent) (Zhao et al., 2015). Amplayo et al. (2018) proposes using translated sentences as context to improve the accuracy of the classifier used for classification tasks. The most recent work in this field, proposed by Shin et al. (2019), presents an embedding distillation framework that significantly decreases the dimensions of word embeddings without compromising accuracy. Several models have been suggested for performing subjectivity analysis (Shen et al., 2018; Zhao et al., 2018; Cer et al., 2018; Radford et al., 2017; Khodak et al., 2018), however the AdaSent model remains the state-of-the-art model in this domain.

2.2 Text Representation Techniques

Text representation refers to the conversion of sequential, textual data into a numeric form so that it can be processed by computer systems that are incapable of dealing with the raw textual data. The text representation approaches range in complexity from simple n-gram (Cavnar et al., 1994) and bag-of-words (Paltoglou and Thelwall, 2013) models to the advanced and efficient ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) models.

Many models utilize deep learning to achieve state-of-the-art results on common NLP tasks, and these models require large amounts of data, training time as well as computational resources to train from scratch. Extensive research has demonstrated that pre-training language models on a large corpus and then fine tuning on task specific datasets can be beneficial for various NLP tasks. (Hube and Fetahu, 2018)

Word2Vec, proposed by Mikolov et al. (2013), is a group of self-supervised shallow two-layer neural network models, that produces word embeddings. Words with similar linguistic contexts are mathematically grouped together in a vector space, which preserves the semantic relationship between words. Vectors extracted from this process are known as word embeddings and these can be used to produce predictions on a word's meaning (Chen and Sokolova, 2018). The major disadvantage associated with this method is that out-of-vocabulary words cannot be represented and it cannot handle polysemous words.

Pennington et al. (2014) presented GloVe (Global Vectors for Word Representations) word embedding which seeks to transform the words into vectors. GloVe is a count based model. The model creates a matrix of co-occurrences of words and performs some dimensionality reduction to learn the word vectors (Agnihotri and Zymbler, 2019). The major drawback associated with this technique is that the vector representations of opposite word pairs, for example, *good* and *bad* are usually located very close to each other in the vector space, limiting the performance of word vectors. Another disadvantage of this model is the fact that representations of out-of-vocabulary words cannot be learned.

Another popular approach is the ELMo (Peters et al., 2018) word representation technique, which considers sub-word units of a word. This model is capable of capturing the meaning and syntactic aspects of words by incorporating internal representations of the network of LSTMs. Since ELMo (Peters et al., 2018) is based on a language model, each token representation is a function of the entire input sentence, enabling it to overcome the limitations of previous word embeddings where each word is usually modeled as an average of its multiple contexts (Perone et al., 2018). The main problem with ELMo is that it is just a shallow concatenation of independently trained left-to-right and right-to-left LSTMs, meaning that the representation cannot take advantage of both left and right contexts simultaneously (Devlin et al., 2019). These problems have been solved by the next language model, BERT.

Devlin et al. (2019) released BERT, the first fine-tuning based representation model that achieves state-of-the-art results on various NLP tasks, making a huge breakthrough in multiple research areas. Trained on a large cross-domain corpus, BERT is designed for two pre-trained tasks: masked language model task and next sentence prediction task. In contrast to ELMo, BERT is not limited to the simple combination of two unidirectional language models. Instead, BERT utilizes a masked language model to predict words which are masked at random to capture bidirectional and contextual information (Devlin et al., 2019). One of the most remarkable features about BERT is that merely fine-tuning the released model can generate significantly good results, especially on small datasets.

3 Experimental Setup

3.1 Dataset Used

We evaluated the various models on the Cornell Subjectivity Dataset v1.0 (SUBJ) (Pang and Lee, 2004).

SUBJ consists of 10,000 records divided into 5,000 subjective sentences extracted from Rotten Tomatoes reviews and 5,000 objective sentences extracted from IMDB plot summaries. Examples of statements from the dataset for both subjective and objective classes are given in Table 1.

Subjective	Objective
If you love motown music, you'll love this documentary. (*opinion*)	'The Journey' is the story of a young Icelandic girl named Kaja. (*fact*)

Table 1: Example of subjective and objective statements from the dataset.

3.2 Embedding Methods

We use two context-independent embedding models, *Word2Vec* and *GloVe*, and two context-dependent embedding models, *ELMo* and *BERT*.

Context-Independent Word2Vec and GloVe have only one numeric representation or embedding for a word regardless of where the word occurs in a sentence or the different meanings it might have. For Word2Vec, we use pre-trained vectors trained on a subset of the Google News dataset (Mikolov et al., 2013). The model contains 3 million words and phrases, represented as 300 dimensional vectors. We use GloVe embeddings (Pennington et al., 2014), pre-Trained on Wikipedia 2014 + Gigaword 5, containing 6 billion uncased tokens. For this work, 50, 100, 200 and 300 dimensional vectors were used but we ultimately settled with the 300 dimensional vectors as they gave the best trade-off between accuracy and computational time.

Context-Dependent ELMo and BERT can generate different word embeddings for a word in order to capture the context of a word in different sentences. ELMo uses two layers of LSTMs to capture the forward and backward information of a word, whereas BERT uses Bidirectional Transformers - an attention based model with positional encodings to represent word positions (Devlin et al., 2019). This model is fairly expensive to pre-train and can be fine-tuned with a few additional layers (Sun et al., 2019). We use ELMo embeddings pre-trained on the 1-Billion Word Benchmark, and the BERT-Base uncased model, consisting of 12 layer transformer blocks, 12 heads, 768 hidden units, and 110 million parameters in total (Devlin et al., 2019).

3.3 Implementation Details

For all the models, we use a Deep Learning pipeline (Figure 1) to carry out subjectivity analysis on the given dataset . We use the pre-trained word embeddings as an Embedding layer. This acts as an interface between the input and the LSTM Layer, which is used for learning long-distance dependencies between word sequences in short texts. The output of the LSTM is connected to a fully connected layer with softmax classifier. While using BERT language model, we consider the final hidden state h of [CLS] token to represent the complete sequence. The probability of x being labelled as class c (subjective) is predicted by a softmax classifier as,

$$P(c|h) = softmax(W^{\mathrm{T}}{}_{\mathrm{classification}} \cdot h) \quad (1)$$

where $W^{\mathrm{T}}{}_{\mathrm{classification}}$ is the weight vector. Training is performed using Adam optimizer. We use binary cross-entropy loss function given by the equation,

$$L(y, \hat{y}) = -\sum{}_{\mathrm{i}}(y_i \cdot log(\hat{y}_i)) \quad (2)$$

where $\hat{y} = P(c|h)$; i.e., the predicted probability for a sentence being subjective while y is the actual class label.

We perform k-fold cross-validation on the dataset as well as the dropout regularization technique to avoid over-fitting. 10% of the entire dataset is considered as validation data in each fold. The rest of the data is used for training. The final accuracy is calculated as the average of accuracy obtained at each fold.

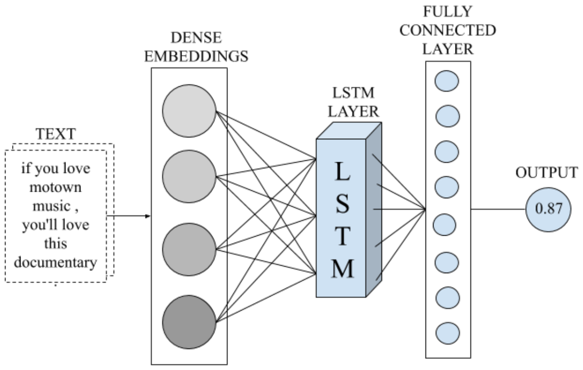
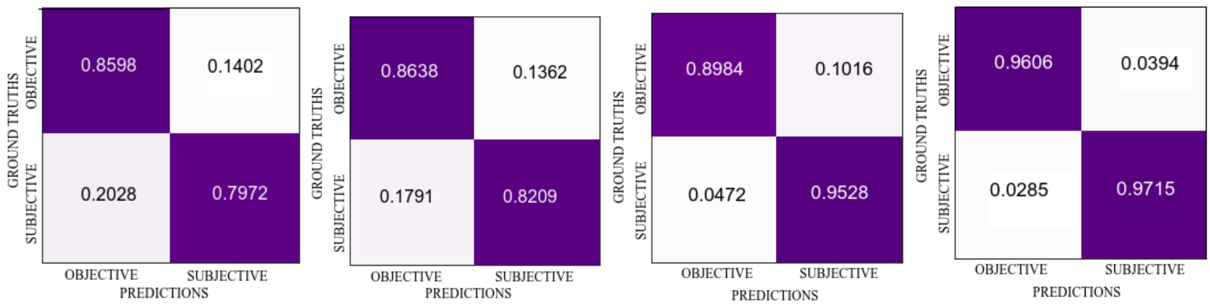


Figure 1: The deep learning pipeline implemented for prediction.

4 Results

The performance of each model is measured by the accuracy, precision, recall and F1-score. The results of the experiments on the SUBJ dataset are presented in Table 2. We present the best performance for each method over the given



(a) Word2Vec (b) GloVe (c) ELMo (d) BERT

Figure 2: Confusion matrices for the four models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
State-of-the-art				
AdaSent	95.50	/	/	/
Context Independent				
Word2Vec + LSTM	82.80	82.50	83.00	83.00
GloVe + LSTM	84.20	84.20	84.50	84.0
Context Dependent				
ELMo + LSTM	91.80	93.00	92.50	93.00
BERT-Base + LSTM	**96.60**	96.40	96.50	96.50

Table 2: Performance(%) of various models.

dataset. Context-independent embedding models, like Word2Vec and GloVe, perform almost similarly on the dataset and are quite far behind the other two models, ELMo and BERT. This is attributed to the fact that these models just take one numeric representation for a word regardless of its relative positioning or multiple meaning. Context-dependent embedding model, ELMo, comes quite close to AdaSent (Zhao et al., 2015), falling short by less than 4% in the accuracy metric. Currently, BERT (Devlin et al., 2019) is one of the most advanced language models, due to its ability to capture bidirectional and contextual information. The BERT-Base model, performs better than AdaSent, achieving an accuracy of 96.60%. The result obtained clearly highlights the importance of context dependency for the subjectivity task.

5 Conclusion

The performance of various language models on the task of subjectivity analysis helps us in drawing important inferences regarding the internal structure of these models and how it corresponds to the results obtained. We compare the performance of each method for the task of Subjectivity Analysis. Our research paves the way to simplify the task of Sentiment Analysis through a Subjectivity Analysis filter which discards the objective statements as they do not offer any opinion. This will lead to faster processing times and less drain on resources due to accurate screening out of objective statements.

References

Sachin Agnihotri and Mikhail Zymbler. 2019. A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, 6.

Reinald Kim Amplayo, Kyungjae Lee, Jinyeong Yeo, and Seung-won Hwang. 2018. Translations as additional contexts for sentence classification. *arXiv preprint arXiv:1806.05516*.

N. Bastas, G. Kalpakis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris. 2019. A comparative study of clustering methods using word embeddings. In *2019 European Intelligence and Security Informatics Conference (EISIC)*, pages 54–61.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Qufei Chen and Marina Sokolova. 2018. Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries. *CoRR*, abs/1805.00352.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1779–1786, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

M. Karamibekr and A. A. Ghorbani. 2013. Sentence subjectivity analysis in social domains. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 268–275.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Georgios Paltoglou and Mike Thelwall. 2013. More than bag-of-words: Sentence-based document representation for sentiment analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 546–552.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Christian S. Perone, Roberto Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444.

Vishnuprakash Sharma, Ajay Panchal, and Vijaya Yogesh Rane. 2020. An analysis on current research trends and applications of natural language processing. *Advance and Innovative Research*, page 63.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *CoRR*, abs/1805.09843.

Bonggun Shin, Hao Yang, and Jinho D. Choi. 2019. The pupil has become the master: Teacher-student model-based word embedding distillation with ensemble learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3439–3445. International Joint Conferences on Artificial Intelligence Organization.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

T. Young, D. Hazarika, S. Poria, and E. Cambria. 2018. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75.

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-adaptive hierarchical sentence model. In *Twenty-fourth international joint conference on artificial intelligence*.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Soufei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. *CoRR*, abs/1804.00538.