TeamCEN at SemEval-2018 Task 1: Global Vectors Representation in Emotion Detection

Anon George, Barathi Ganesh HB, Anand Kumar M and Soman KP
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
anongeorge007@gmail.com, barathiganesh.hb@gmail.com

Abstract

Emotions are a way of expressing human sentiments. In the modern era, social media is a platform where we convey our emotions. These emotions can be joy, anger, sadness and fear. Understanding the emotions from the written sentences is an interesting part in knowing about the writer. In the amount of digital language shared through social media, a considerable amount of data reflects the sentiment or emotion towards some product, person and organization. Since these texts are from users with diverse social aspects, these texts can be used to enrich the application related to the business intelligence. More than the sentiment, identification of intensity of the sentiment will enrich the performance of the end application. In this paper we experimented the intensity prediction as a text classification problem that evaluates the distributed representation text using aggregated sum and dimensionality reduction of the glove vectors of the words present in the respective texts.

1 Introduction

Emotion detection from text has been an important task in recent years since the development of Natural Language Processing. Finding the affect in tweet enlarges the business intelligence towards the consumer behaviour analysis, peoples likeness towards the person, organization and policies of the government.

Intensity of the sentiment present in the text can be predicted by performing a text classification by taking texts as the observations and their intensity classes or scores as a target labels. Representation is a key part in any text classification task which can affect the further feature extractions and predictions (Ganesh et al., 2016; Soman et al., 2016; B. et al., 2016). A bad representation of the word vector can make the further predictions fruitless. Hence we use a meaningful representation of words, that is, global vectors (Glove) for the representation of words into vectors of a tweet. This paper explains the use of global vectors representation for representing the words in a tweet and their further processing using machine learning techniques for classification and regression tasks. SemEval-2018 Task 1 provided the twitter corpus for classification and regression tasks for the emotion detection.

After representing the words in the tweets, the vector for tweets is obtained by computing the aggregated sum and dimensionality reduction. The final tweet vectors are used for further classification and regression.

2 Global Vectors

Global Vectors (GloVe) creates word vectors and it is an unsupervised machine learning technique(Pennington et al., 2014). Unlike word2vec representations like skip-gram and continuous bag of words (CBOW), word-word co occurrence statics is taken for the vector representation. It also retains the relations between words and gives a better feature extraction. The vectors represented in space gives more meanings than the traditional methods. global matrix factorization methods and local context window methods, such as the skip-gram model (Mikolov et al., 2013) is the main two types of vector representations. The word-word co occurrence count on the whole set of words are used for training and hence it captures more information.

3 Datasets

Dataset consists of tweets from three languages that are English, Spanish and Arabic. These are mainly focused on the emotions contained in it. These are annotated for the different emotions

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 334–338
New Orleans, Louisiana, June 5–6, 2018.

such as joy, fear, anger and sadness. Separate datasets are provided for each emotions. The classification tasks consists of annotations marked as different intensity values for each emotion. They also contains the value corresponding to how much emotion value can be inferred from the tweets. For the regression tasks, the values between 0 and 1 are provided along with it. The value 0 means no information can be inferred and 1 means maximum information can be inferred. Sub-tasks EI-oc, V-oc, E-c are multi-class classification problems. EI-reg and V-reg are regression problems.EI-reg and EI-oc are emotion intensity tasks and V-reg is a sentiment intensity task. V-oc is a sentimental analysis task and E-c is a emotion classification task. (Mohammad and Kiritchenko, 2018)

4 Methodology

Pre-trained model of GloVe with 100 dimensions computed from 27 billion tokens and 2 billion tweets from the twitter is used in this experiment [1]. From the tweet datasets, the words are taken and checked if it is present in the downloaded GloVe model. If it is present then the vector representation corresponding to that word is taken and added to the model. After obtaining the vector representation for all the words in a tweet, the vector representation for each tweet is made after pre-processing steps like unimportant symbols and space removal. The representation for a tweet is made by writing the word vector as columns of a matrix concatenated till the tweet ends. This matrix is reduced into a single vector by two methods:

- SUM: Taking the sum of each rows and making it a new vector. This is an easy method which has less computation but the disadvantage includes the loss of word order in a tweet.
- SVD: Singular Value Decomposition is a decomposition method widely used for reducing the dimension of matrices. Here the sentence matrix is reduced to a rank one matrix by taking only the most important singular value. This is then used to take the important vector from the matrix which can be the most information containing (Golub and Reinsch, 1970). The tweet matrix is decomposed 3 parts using SVD. The most important singular value occurs in the first value of the diagonal matrix. A new vector is formed by using only one singular value from the SVD. Reducing the dimension using SVD preserves the important spread of data rather than simple summing operations of columns.

$$A = U \sum V^T \tag{1}$$

These vectors are used as the input for the different machine learning techniques. A part of the training data is split into training and validation data and is given to learning methods like Random Forest and Support Vector Machine (SVM). These methods will learn and give predictions about the classification or regression task that is assigned. Experiments are done to find out which hyper parameters give the better value for the validation accuracy. These are stored and used for predicting the new unseen data from the test set.

4.1 Random Forest

Random Forest is a powerful and popular machine learning algorithm capable of performing both regression and classification tasks. This algorithm creates a number of decision trees. When the number of these decision trees are more , the predictions will have more robustness and accuracy. Random Forest consists of multiple trees to classify a new object based on attributes. A classification is given by each tree and the tree votes for that class are saved. Choice is made for the classification by taking the most votes for consideration. It can handle the missing values and can maintain accuracy for the missing data. It can also work on large datasets which are high dimensional. It is not that good at regression like it does for the classification tasks. The control over the model is less as there are less parameters to tune.

Test features are passed through the rules of each randomly created trees for the predictions. Each tree will be predicting an output and the voting is done for the prediction to get the highest vote for a particular class of prediction. This is called majority voting. It is called ensemble machine learning algorithm. Divide and conquer approach is used for the improvement of performance. Each classifier in this group may be weak learner, but when they come together, it acts as a strong learner (Liaw et al., 2002).

[1] https://nlp.stanford.edu/projects/glove/

4.2 Support Vector Machine (SVM)

SVM is used to classify different classes in a dataset. It draws a decision boundary after looking at the extreme data points in a dataset. This boundary is called the hyper-plane which has one dimension less than the dimension of the data points. It is drawn near the extreme points of the dataset. SVM is a popular algorithm which segregates the different classes of data points in a dataset. If the decision boundary is drawn without making it optimized, then the further classifications will have less accuracy on new data. Support Vectors are the data points that are closer to the other classes and they are the ones pushing the boundary farther to make better predictions. This algorithm says that only those support vectors or the margins are needed for the further classifications and the other data points are ignored. This is because the margin is drawn by considering the extreme case in a class and all the other points can easily be classified with that prior knowledge.

SVMs can also be used in higher dimensional datasets and the data points will be called as vectors and they will have their coordinates lying inside the space of data. For the cases in which our function is not linearly separable, our data should be transformed in to a higher dimension using a function. Data points in higher dimensions are computationally complex for predictions. Kernel trick can be used to reduce this computational cost. A kernel function or a kernel trick is a function that takes the input vectors from the original vector space and returns the dot product of the vectors in the features in the feature space. To map every point into a higher dimensional space, dot product between two vectors can be applied using some transformation. Transformation on non-linear space into linear space is possible using that (Suykens and Vandewalle, 1999). Common kernel types are,

1. Linear Kernel
2. Polynomial Kernel
3. Radial Basis Function (RBF) Kernel

Table 1 shows the kernel functions of different kernel. Here x and y are the data points and the kernel function is found using the equations corresponding to it. d is the dimension of the space and γ is a hyper parameter. To get better performance using any kernel, parameter tuning is required. SVMs are effective in higher dimensions and it is possible to add custom kernels to it making it adaptive. It performs poor when the number of features are greater than the number of samples.

Kernel Name	Kernel Function
Linear Kernel	$k(x, y) = x \times y$
Polynomial Kernel	$k(x, y) = (x \times y + 1)^d$
RBF Kernel	$k(x, y) = e^{-\gamma\|x-y\|^2}$

Table 1: Kenel Functions

We have reported the observations made using the linear and RBF kernel.

5 Results and Observation

Mean Squared Error (MSE) is noted for all the regression tasks and accuracy is noted for all the classification tasks. MSE near the value 0 is always better and accuracy value can be a maximum of 100%. Accuracy measures are in percentage. Validation scores for the test data provided is measured and reported for observations.

Using the SVD method always gave slightly better result compared to the SUM method. From this we can understand that the the reduction of a matrix in to vector should be done with SVD instead of SUM method. Scores for the language English gave better results compared to Spanish and Arabic. This can be due to the use of pre-trained GloVe vectors which was trained on more English data.

Table 2 has MSE for validation set in EI-reg (emotion intensity regression). For English, Random Forest gave slightly better results than the other two SVM methods. Random Forest and SVM with RBF kernel were performing nearly same for the Spanish and Arabic datasets.

Table 3 has Accuracy of validation set in EI-oc (emotion intensity ord.class.). Random Forest Classifier was performing better than the other two classifiers for all the three languages. English language showed better accuracy than the other two languages and Arabic showed least accuracy.

Table 4 has MSE for validation set in V-reg (valence intensity regression). All the three regressors were performing in a similar way for English wereas Random Forest and the SVM with RBF kernel performed better for the other languages.

Table 5 contains Accuracy for validation set in V-oc (valence ord.class.). Random Forest was performing better for all the classification tasks

	Random Forest		SVM (Linear)		SVM (RBF)	
	SUM	SVD	SUM	SVD	SUM	SVD
English	0.03	0.02	0.03	0.03	0.04	0.03
Spanish	0.04	0.03	0.05	0.05	0.04	0.03
Arabic	0.03	0.02	0.03	0.03	0.03	0.02

Table 2: Mean Squared Error for validation set in EI-reg (emotion intensity regression)

	Random Forest		SVM (Linear)		SVM (RBF)	
	SUM	SVD	SUM	SVD	SUM	SVD
English	45.42	47.31	43.36	45.28	42.55	44.45
Spanish	40.35	42.20	39.72	42.12	38.58	40.56
Arabic	28.74	31.05	26.34	28.03	25.06	26.98

Table 3: Accuracy for validation set in EI-oc (emotion intensity ord.class.)

	Random Forest		SVM (Linear)		SVM (RBF)	
	SUM	SVD	SUM	SVD	SUM	SVD
English	0.05	0.04	0.05	0.04	0.05	0.04
Spanish	0.04	0.03	0.04	0.04	0.04	0.03
Arabic	0.05	0.05	0.06	0.05	0.06	0.05

Table 4: Mean Squared Error for validation set in V-reg (valence intensity regression)

	Random Forest		SVM (Linear)		SVM (RBF)	
	SUM	SVD	SUM	SVD	SUM	SVD
English	26.05	28.66	22.49	25.04	22.01	23.54
Spanish	23.58	24.05	19.65	20.12	18.05	19.57
Arabic	20.28	22.34	13.04	13.64	12.50	13.62

Table 5: Accuracy for validation set in V-oc (valence ord.class.)

	Random Forest		SVM (Linear)		SVM (RBF)	
	SUM	SVD	SUM	SVD	SUM	SVD
English	95.43	95.56	95.14	95.68	94.56	94.66
Spanish	95.14	95.41	95.43	95.54	95.01	95.34
Arabic	93.84	93.91	93.84	94.02	92.12	92.31

Table 6: Accuracy for validation set in E-c (multi-label emotion class.)

	Pearson (all instances)				
	macro-avg	anger	fear	joy	sadness
English	0.077 (44)	0.062 (44)	0.076 (44)	0.079 (43)	0.090 (44)
Arabic	0.230 (10)	0.213 (10)	0.230 (10)	0.207 (11)	0.269 (11)
Spanish	0.131 (12)	0.184 (11)	0.117 (12)	0.143 (11)	0.078 (12)

Table 7: Published Results score for EI-reg where the number in brackets is the ranking

given. Accuracy for Arabic stayed down and English stayed up. Table 6 contains Accuracy for validation set in E-c (multi-label emotion class.). All the classifiers for all languages were performing nearly same for this task. English still remained on top where as SVD always gave a slight boost in accuracy like the previous cases.

The published results for all these are also tabulated. The algorithm did not give the score and accuracy that was giving on the validation sets.

Table 7 contains published Results score for EI-reg where the number in brackets is the ranking. The minimum value is 0 and maximum value is 1. Our algorithm has lower score but comparatively the sentiment "sadness" has slightly better score. Score of published results for EI-oc was not satisfactory and the score values were less. In published results score for V-reg, Valence score for Arabic language (0.319) was better than the other 2 languages and English had the lowest. For the published results score for V-oc, algorithm's performance was not satisfactory but the arabic language had better valence (0.163) than the other 2 languages. In the published results score for E-c, the accuracy for Spanish was slightly better compared to the English and Arabic languages.

6 Conclusion

Five different tasks with common framework is experimented to find out the affect in tweets in 3 different languages. From the scores and accuracy obtained from the validation data we can conclude that the Global Vector Representation gives good results for the sentiment analysis tasks. Using a pre-trained model for the GloVe made the task simpler and easy to use for vector representation. Reducing the tweet matrix into vector using SVD always gave better results than taking the column wise sum of the matrices. This shows the importance of the word order. Random Forest algorithms gave better results for the classification tasks were as the SVD algorithm with RBF kernel performed nearly well in the regression tasks. English language showed better scores in the validation data compared to other languages. From these observations, it can be noted that this approach performed satisfactorily better in the validation step and was able to get the semantic features from the tweets. Hence the future work will be focused more on the performance of the experimented methods.

References

Barathi Ganesh H. B., M. Anand Kumar, and K. P. Soman. 2016. Statistical semantics in context space : Amrita_cen@author profiling. In *CLEF*.

HB Barathi Ganesh, M Anand Kumar, and KP Soman. 2016. From vector space models to vector space models of semantics. In *Forum for Information Retrieval Evaluation*, pages 50–60. Springer.

Gene H Golub and Christian Reinsch. 1970. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Jeffrey Pennington, Richard Socher, and Christopher 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.

KP Soman et al. 2016. Amrita_cen at semeval-2016 task 1: Semantic relation from word embeddings in higher dimension. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 706–711.

Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.