Amrita-CEN-SentiDB:Twitter Dataset for Sentimental Analysis and Application of Classical Machine Learning and Deep Learning

Authors:
K S Naveenkumar
R Vinayakumar
K P Soman

Department of Computational Engineering and Networking

Amrita School of Engineering

Amrita Vishwa Vidyapeetham

16-MAY-2019

Outline of the Presentation

- Introduction
- Motivation
- Objective
- Description of the Data set Collected
- Implemented Architecture
- Features that are extracted
- Visualization of the data set
- Methodology
- Results
- Conclusion
- Future Works

Introduction

- Natural language processing is an area of computer science and artificial intelligence which are concerned with the interactions between the computers and human languages.
- Sentimental Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Introduction(contd..)

- Sentiments are the combinations of the feelings, behavior, physiology, conceptualization and experience [1][2] that are expressed by any living beings.
- Each human has their identity in the digital world through the social medias such as the Facebook, Twitter, Instagram, Gmail, Snapchat, Whatsup and what so ever application the user is bound to use it. Tons of the text are generated each and every day in these social media.

Motivation

- All the decisions are taken based on the emotional stability of the person at that instance.
- My aim is to create a trained model which classifies the sentiment (Positive, Negative or Neutral) from the given text data.
- Sentiment based unstable decisions by the people can be prevented once the system is introduced. Example: Blue whale game.

Objective

- To create a trained model that would predict the emotion (positive,negative and neutral) of the person from the text with the dataset that has been collected and trained with the various data driven models.

Description of the Data Set Collected

- Twitter dataset from available sources
- Training details
  - All these datasets have been combined to get a total of 1,53,642 sentences in which 70% of them are taken for training such that 1,07,550 sentences are taken for training.
- Test details
  - For the test data 30% is split comes to a count of 46,092 sentences.
- The collected dataset is made publically available for research purpose in the link [1]

Category	**Sentences**
Train data	1,07,550
Test data	46,092
Positive	62,629
Negative	55,477
Neutral	35,536

Table 1: Dataset Details

[1] https://vinayakumarr.github.io/Amrita-CEN-SentiDB/

Implemented Architecture

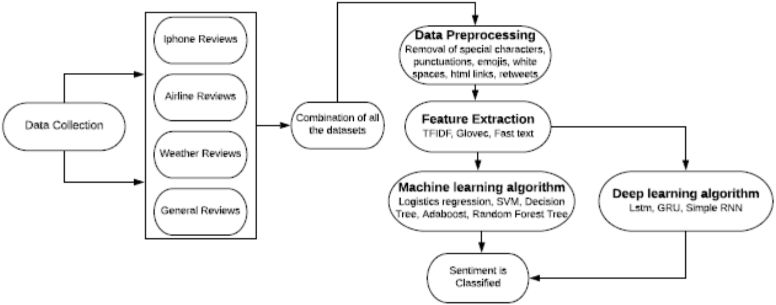
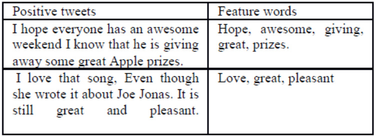

Data Collection
Iphone Reviews
Airline Reviews
Weather Reviews
General Reviews
Combination of all the datasets
Data Preprocessing
Removal of special characters, punctuations, emojis, white spaces, html links, retweets
Feature Extraction
TFIDF, Glovec, Fast text
Machine learning algorithm
Logistics regression, SVM, Decision Tree, Adaboost, Random Forest Tree
Deep learning algorithm
Lstm, GRU, Simple RNN
Sentiment is Classified

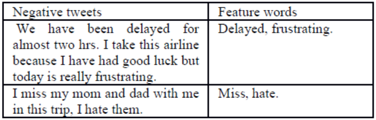
Figure 1: Architecture Diagram

Features that are extracted



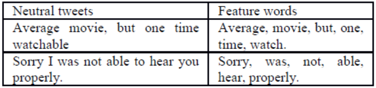
Positive tweets	Feature words
I hope everyone has an awesome weekend I know that he is giving away some great Apple prizes.	Hope, awesome, giving, great, prizes.
I love that song, Even though she wrote it about Joe Jonas. It is still great and pleasant.	Love, great, pleasant

Figure 2: Features from positive sentence



Negative tweets	Feature words
We have been delayed for almost two hrs. I take this airline because I have had good luck but today is really frustrating.	Delayed, frustrating.
I miss my mom and dad with me in this trip, I hate them.	Miss, hate.

Figure 3: Features from negative sentence



Neutral tweets	Feature words
Average movie, but one time watchable	Average, movie, but, one, time, watch.
Sorry I was not able to hear you properly.	Sorry, was, not, able, hear, properly.

Figure 4: Features from neutral sentence

Visualization of the Data Set

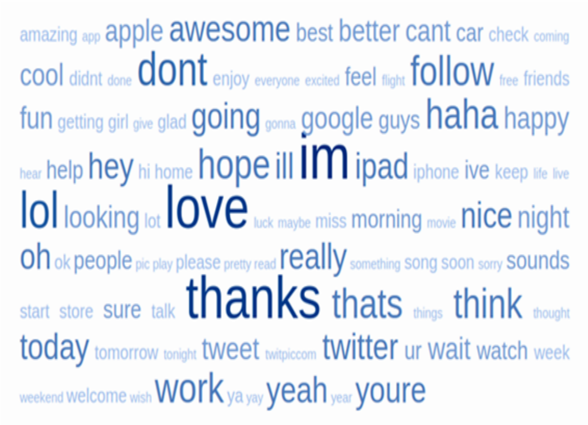


Figure 5: Plot for Positive Data set

Visualization of the Data Set

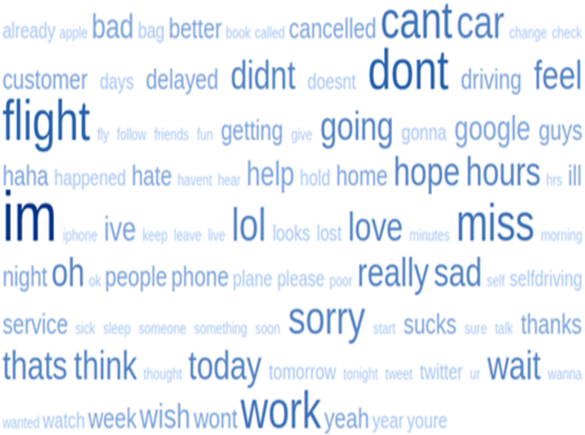


Figure 6: Plot for Negative Data set

Visualization of the Data Set

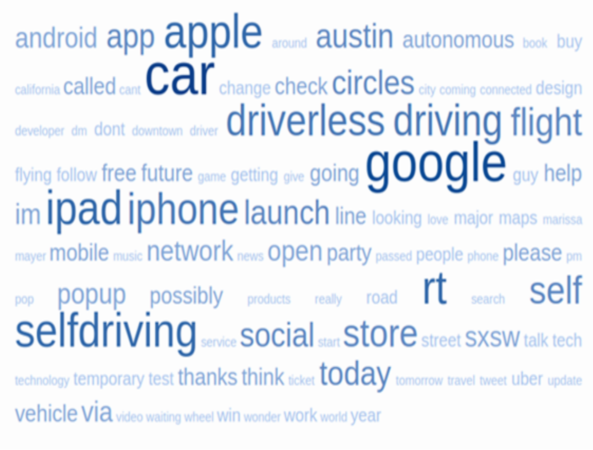


Figure 7: Plot for Neutral Data set

Methodology

- Machine Learning
  - Feature extraction : TFIDF (Term Frequency Inverse Document Frequency) and then to the classifiers like Random forest tree, Decision tree, SVM-Liner and Rbf, Adaboost and Logistics Regression.
- Deep Learning
  - Using Keras embedding techiques for the word level feature extraction and then passed to the LSTM (Long Short Term Memory) and then CNN (Convolution Neural Network)

Results and discussions

Features	**Classifiers**	**Accuracy**	**Precision**	**Recall**	**F-score**
10000	Decision tree	0.642	0.642	0.642	0.642
	Adaboost	0.652	0.655	0.652	0.648
	Randomforest tree	0.727	0.727	0.727	0.727
	SVM Linear	0.751	0.750	0.751	0.750
	SVM rbf	0.476	0.226	0.476	0.307
	Logistics regression	**0.753**	**0.753**	**0.753**	**0.752**
20000	Decision tree	0.642	0.642	0.643	0.643
	Adaboost	0.650	0.650	0.650	0.648
	Randomforest tree	0.727	0.727	0.727	0.727
	SVM Linear	0.755	0.755	0.755	0.755
	SVM rbf	0.476	0.226	0.476	0.307
	Logistics regression	**0.756**	**0.756**	**0.756**	**0.756**
30000	Decision tree	0.647	0.646	0.647	0.647
	Adaboost	0.657	0.657	0.657	0.655
	Randomforest tree	0.728	0.728	0.728	0.728
	SVM Linear	**0.757**	**0.757**	**0.757**	**0.757**
	SVM rbf	0.476	0.226	0.476	0.307
	Logistics regression	0.756	0.756	0.756	0.755
40000	Decision tree	0.647	0.647	0.647	0.647
	Adaboost	0.653	0.653	0.653	0.651
	Randomforest tree	0.729	0.728	0.729	0.728
	SVM Linear	**0.758**	**0.758**	**0.758**	**0.757**
	SVM rbf	0.476	0.226	0.476	0.307
	Logistics regression	0.757	0.757	0.757	0.756

Table 2: Results from Classical Machine Learning

Results and discussions

Features	Algorithm	Accuracy	Precision	Recall	F-score	Time for computing
Keras Embedding	LSTM	0.447047	0.447047	0.447281	0.447032	1680 minutes
	CNN	**0.452596**	**0.452596**	**0.446198**	**0.446877**	**1440 minutes**

Table 3: Results of Deep Learning

Conclusion

- The paper evaluates the performance of linear and non-linear text representation methods for sentimental analysis.
- The collected dataset Amrita-CEN-SentiDB is subjected to various non-linear text representation methods with the deep learning architecture which performs better than the linear text representation with the machine learning algorithms.

Future Works

- The performance of the proposed method can be increased experimentally by hyper parameter tuning the network. This is the benchmark accuracy for this dataset further the dataset is made publically available for the research purpose.

References

[1] Mohammad, Saif M., and Felipe Bravo-Marquez. "WASSA-2017 shared task on emotion intensity." arXiv preprint arXiv:1708.03700 (2017).

[2] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Evaluating deep learning approaches to characterize and classify malicious URLs." Journal of Intelligent Fuzzy Systems 34.3 (2018): 1333-1343.

[3] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." Procedia Computer Science 17 (2013): 26-32.

[4] Mohammad, Saif, et al. "Semeval-2018 task 1: Affect in tweets." Proceedings of The 12th International Workshop on Semantic Evaluation. 2018.

[5] Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis. "Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis." Proceedings of the 11th International Workshop in SemEval-2017.

Thank You . . .