Amrita-CEN-SentiDB1:Improved Twitter Dataset for Sentimental Analysis and Application of Deep learning

Authors:
K S Naveenkumar
R Vinayakumar
K P Soman

Department of Computational Engineering and Networking

Amrita School of Engineering

Amrita Vishwa Vidyapeetham

07-JULY-2019

Outline of the Presentation

- Introduction
- Objective
- Description of the Data set Collected
- Implemented Architecture
- Features that are extracted
- 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.

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.

Objective

The main objectives of this work are as follows:

- To develop a Twitter database for the sentimental analysis.
- To perform various text representation methods and to evaluate those methods on the sentimental database.
- To do a comparative study on the sentimental database by using deep learning techniques.

Description of the Data Set Collected

- Twitter dataset from available sources.
- The collected dataset is made publically available for research purpose in the link. [1]

Dataset	Split	Positive	Negative	Total
Training	60:40	1,20,000	1,20,000	4,00,000
Testing		80,000	80,000	
Training	70:30	1,40,000	1,40,000	4,00,000
Testing		60,000	60,000	
Training	80:20	1,60,000	1,60,000	4,00,000
Testing		40,000	40,000	
Training	90:10	1,80,000	1,80,000	4,00,000
Testing		20,000	20,000	

Table 1: Description of the Amrita-CEN-SentiDB1 dataset

- Here the 70:30 split gives the best performance, results obtained are tabulated below.

[1] https://vinayakumarr.github.io/Amrita-CEN-SentiDB1/

Implemented Architecture

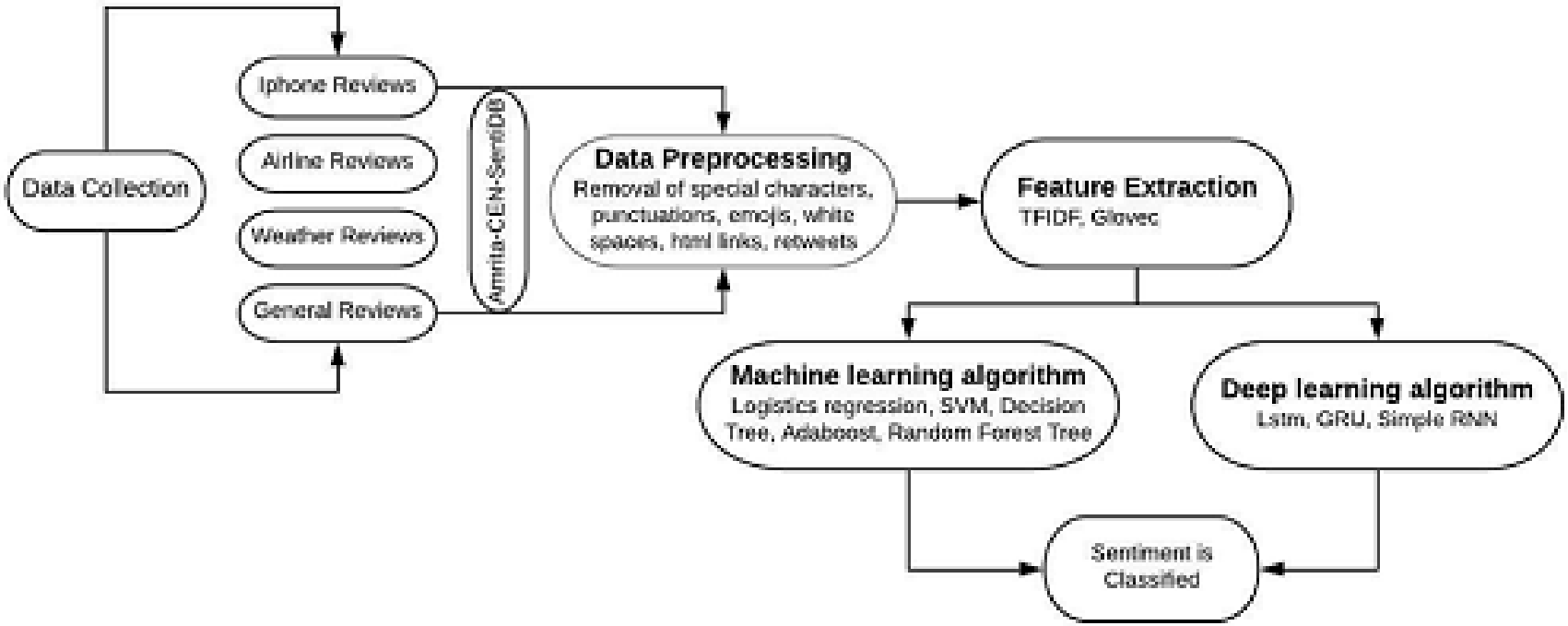

Figure 1: Architecture Diagram

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

Methodology

- Deep Learning
  - TFIDF (Term Frequency Inverse Document Frequency) and Glovec then to the classifiers like LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit), SimpleRNN (Recurrent Neural Network)

Results and discussions

Features	Classifiers	Accuracy	Precision	Recall	F-Score
10,000	LSTM	49.2	53.2	42.6	44.9
20,000		50.3	42.6	44.5	43.8
30,000		49.1	33.9	32.5	33.2
40,000		50.8	50.1	50.3	50.3
10,000	GRU	53.5	51.2	51.6	50.7
20,000		49.2	48.6	48.7	47.4
30,000		55.3	52.5	53.8	48.6
40,000		51.2	50.5	50.1	50.2
10,000	Simple RNN	51.1	52.3	52.9	52.9
20,000		50.5	50.6	49.6	41.1
30,000		50.6	49.8	41.8	50.2
40,000		53.9	50.5	49.2	52.3

Table 2: RESULTS USING DEEP LEARNING BY TFIDF APPROACH

Results and discussions

Method	Classifiers	Accuracy
Glovec	LSTM	75.3
	GRU	63.2
	Simple RNN	69.5

Table 3: RESULTS USING DEEP LEARNING BY GLOVEC APPROACH

Conclusion

- The paper evaluates the performance of linear and non-linear text representation methods for sentimental analysis.
- The collected dataset Amrita-CEN-SentiDB1 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 ...