

General Depiction of an Efficient Bootstrapper: A Review

Rabiya Qadir*, Afaq Alam Khan

Department of Information Technology, Central University of Kashmir, Ganderbal,
Jammu and Kashmir, India

Abstract

Bootstrapping is a technique which can automatically annotate emotional corpora. It is a pre-eminent job aimed by many companies who want to research the sentiments of products before purchase, or many companies that want to observe a view or opinion that is held or expressed by public about their brands. The significance of this research is expressed by the fact that, in the present era, people's emotions has a very great importance for business, politics and for people also. So, there arises a need to create a technique which can create emotional corpus and using this resource emotion detection, systems can work in the best possible manner with the least waste of time and effort. In this paper, different techniques and approaches are discussed given by different researchers to create an efficient bootstrapper for automatic annotation of data. This paper also presents the results that were recorded in the different experiments by various researchers.

Keywords: Bootstrapping, emotional corpora, emotion detection system, NLP (natural language processing), SVM (support vector machine)

***Author for Correspondence** E-mail: rabiyaqadri22@gmail.com

INTRODUCTION

In the present era of world with such a high population, data are present in abundance and mining of such data is a very difficult task and is very important for business, politics, etc. As people express their thoughts on social media, and these thoughts play an important role in business, politics, etc. Hence, an analysis of this type of data should be performed and is called sentiment analysis. It is actually mining of data which recognizes and takes out subjective information. This information helps companies or business to understand the social sentiment about their product and in politics it helps Politian's to know about the public opinion.

With the new advances in deep learning, the ability of algorithms has improved substantially. There has been enormous research in the area of sentiment analysis. In order to get sentiments from data, we need to analyze and categorize it. This process of categorization is done by classifiers. These classifiers are first trained. In order to train a classifier supervised learning usually needs hand labelled training data. But this process of manual labelling is time consuming and biased. Therefore, an automatic approach was

introduced which is called bootstrapping. It is the process of automatic annotation of data to create emotional corpora. The emotional corpus created by this approach makes emotion detection systems more efficient and robust.

For years, different methods like bootstrapping is used by many researchers to annotate twitter data, text messages, emoticons, emoji, blogs, etc. Due to the limitation, traditional ways of manual labelling of data were time consuming, costly and biased. Thus automated systems are needed. Trying to sort out the sentiments of people correctly and the high cost issues, data from social media can be used as a resource as it provides enough data source for analyzing important topics and elated public reactions. It also allows the users to express their sentiments or emotion as much as possible.

In this paper, we are trying to succinct various related state of art methods and techniques used in different but related studies together into a single location.

RELATED WORK

An emotional corpus is a database of huge and organized set of sentences where every

sentence is labelled with one or more emotional labels. As supervised learning approaches are dependent on annotated training data which are actually set of examples so these emotional corpora becomes the necessary base for supervised approaches. The supervised learning algorithm examines the training data and deduces an outcome, from which new examples can be mapped.

The establishment of emotional corpora becomes pre-eminent because of the fact that in the area of emotion detection supervised learning method is applied in various approaches.

The machine learning algorithms learn from human annotations as the labelling process of emotional corpora is done manually. In this section, we review the related work regarding bootstrapping technique. Different methods are used by researchers for creating an efficient bootstrapper.

Canals et al. [1] used bootstrapping method for automatic annotation of emotions in texts. They have used Aman corpus as their data set, contained 4000 sentences that were taken from blog posts which were collected directly from web. They chose this corpus because sentences were manually annotated and this will allow researchers to compare and also check the efficacy of the method.

In this approach, they have simply selected a sentence which has an emotional vector associated that contains a value for every emotion. This sentence is then tokenized and lemmatized through natural language processing and after that word is matched with the Emolex, and if word is present in Emolex then its corresponding value is added to it. The sentence is then labelled with the same emotion which bares the highest value in the sentence. Moreover, they have also presented two enhanced approaches for creating a seed and have also enlarged Emolex.

The results obtained from this approach seem promising, and this concludes that a better emotion model can be created using bootstrapping approach. Thus, bootstrapping approach can be considered as fast, easy and a good approach for automatic annotation of data.

Go et al. used a new approach to classify the sentiments of the messages collected from twitter automatically. They have used two categories, which are “positive” or “negative” to classify the messages from twitter with reference to the query term. They have proposed this approach so that people who want to buy a product will know about the view or opinion that is held or expressed by the people about that particular product and also there was no existing research on the sentiment of messages from twitter.

They have used twitter messages as their data set that contains emoticons also which were employed as noisy labels. They used this type of training data because it was easily available in an abundant form and this type of data can also be acquired via automated means. They have used machine learning algorithms and the results acquired by those had accuracy above 80%. They have also narrated the preprocessing steps that are required to attain greater accuracy.

The salient offering of this paper is the proposal of employing tweets that contain emoticons for distant supervision. They have also showed that various feature extractors and machine learning classifiers like SVM, Naïve Bayes, etc. can be used to achieve greater accuracy. These classifiers can also be used for the classification of twitter messages with same performance.

Keshtkar et al. proposed a unique approach to extricate paraphrases for terms which contains emotions. They have used bootstrapping process so that it can learn extraction patterns for emotion classes. They have used this bootstrapping process to recognize paraphrases. This identification process of paraphrasing gets initiated by taking some seeds. These seeds are basically words extracted from word net. They have used labelled blogs and some other data sets like texts to extricate paraphrases based on highest gaining extraction patterns. They have used k-window algorithm to extricate aspirant patterns from seeds. They have included morpho-syntactic and lexical paraphrases in results that have been evaluated by human judges.

Qadir et al. presented a weakly supervised technique based on five emotions for learning hashtags, patterns and phrases. They have used

a bootstrapping framework to label set of tweets through this annotation; they have trained their classifiers and have utilized them to learn new hashtag patterns. They have showed a considerable improvement in f-scores. They have used concept of common prefix to categorize the emotions present in the sentence. They have used tweets as their data set. They have combined emotion indicators with N-Gram classifier which showed a considerable improvement over emotion classes.

Sutles et al. have proposed a technique which can recognize emotions in tweets. They have used eight emotions defined by plutchick as four pairs of opposing emotions which produces a feasible base for binary classification of data taken from twitter. They have taken hash tags, emoticons and have included emoji as well. They have showed that emoji can stipulate more emotions than emoticons. As they have used emoji which can be noisy, for this, they first conduct an experiment to explore the similarity among specific labels of different types which is supposed to be indicative of the same option they have used distant supervision to produce accurate classifiers. They have then tested and compared the correctness of classifiers for the four binary pairs of plutchick emotions. The results produced by the classifiers are between 75% and 91% which actually depends upon the pair of emotion. They have showed that the classifier can be merged to imitate the multilabel classifier for emotions defined by plutchick which have showed greater accuracies than previous ones.

Charbonneau et al. have studied the behavior of individuals in the stressful circumstances which

explained the relationship between the gender and symptoms of depression. They have also hypothesized that this relationship gets impacted by the individual dissimilarity in emotional response. They have explained that their findings hold up the building agreement that there are various alleys that leads to gender dissimilarity in depression. They have explained that the research on particular alley can be utilized to generate certain intercessions. They have explained that the character like withdraw negativity can be taken as a principle element in the developmental course of depression in females. They have also showed that many tools can be included in the screening tools. Through this early incident of depression can be prevented in females.

The general procedure for bootstrapping process can be shown as in Figure 1.

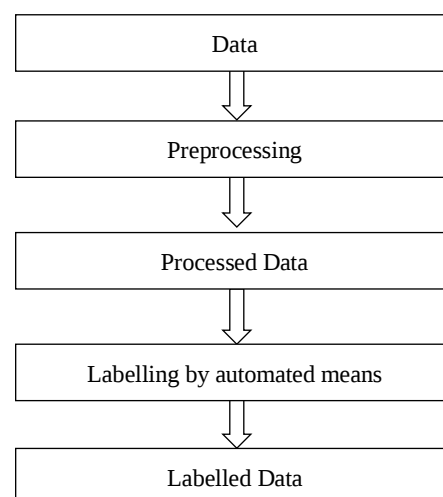


Fig. 1: Overview of Bootstrapping Process.

Summarization of the previous work done is given in Table 1.

Table 1: Summarization of the Results Obtained.

Prior work	Year	Dataset	Approaches used	Results
Canals et al. [1]	2016	Aman Corpus	SVM (support vector machines)	F1=0.689 (Joy)
Go et al. [2]	2009	Data from Twitter	Classifiers: Naïve Bayes, maximum entropy, SVM	Accuracy=82.2% (unigrams, SVM)
Keshtkar et al. [3]	2013	Blogs, text affect data, fairy tales, annotated blog data	Stanford POS (part-of-speech) tagger and chunker, K-window	Recall=92.91% (disgust)
Qadir et al. [4]	2013	Data from Twitter	Logistic regression, emotion N-Gram classifier	The approach used produced better results than previous methods
Sutles et al. [5]	2003	Data from Twitter	Naïve Bayes, maximum entropy, natural language tool kit, cross validation	75-91%
Charbonneau et al. [6]	2009		Multiple regression, constrained non-Linear regression	The approach used added more knowledge and support to the previous studies

CONCLUSION

In this paper, we have discussed how different data like twitter messages, emoticons, texts, etc. can be annotated by using bootstrapping process so that this will lead to create an efficient emotion detection system. This field is important for business, politics and other purposes. It helps in categorizing data based on emotions which will help companies or Politian's to take better decisions about their products or strategies. In this paper, we have given a brief summary of important approaches and techniques which are used in bootstrapping process and can be used to make a large corpus of emotional data. A brief summary of results obtained and datasets used by different researchers are also outlined in Table 1.

REFERENCES

1. Canales L, Strapparava C, Boldrini E, Martnez-Barco P. Exploiting a bootstrapping approach for automatic annotation of emotions in texts. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2016 Oct 17*, 726–734p.
2. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford. 2009 Dec; 1(12).
3. Keshtkar F, Inkpen D. A corpus-based method for extracting paraphrases of emotion terms. *Proceedings of the NAACL HLT 2010 Workshop on Computational approaches to Analysis and Generation of emotion in Text*, Association for Computational Linguistics, 2010 Jun 5, 35–44p.
4. Qadir A, Riloff E. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1203–1209p.
5. Suttles J, Ide N. Distant supervision for emotion classification with discrete binary values. *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, 2013 Mar 24, 121–136p.
6. Charbonneau AM, Mezulis AH, Hyde JS. Stress and emotional reactivity as explanations for gender differences in adolescents' depressive symptoms. *J Youth Adolescence*. 2009 Sep 1; 38(8): 1050–8p.

Cite this Article

Rabiya Qadir, Afaq Alam Khan. General Depiction of an Efficient Bootstrapper: A Review. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(3): 9–12p.