IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 7, Issue 2, February 2019

Survey on Sentiment Analysis

Sheba Jebakani[1], Sujith Surendranath[2], Vinay Kumar B U[3], K R Nagendra[4]

Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[1]

Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[2,3,4]

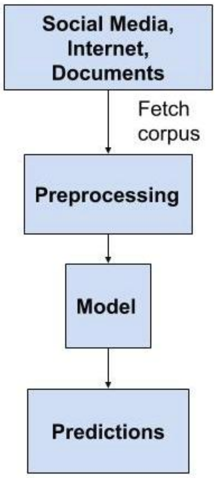
Abstract: Sentiment analysis fundamentally implies the way toward recognizing and characterizing an assessment or explanation in light of the extremity communicated in it particularly to decide the sentiments communicated by the reviewer. In this new period of the web with consistently developing information as the content, we require an approach to group the content into positive, negative and once in a while unbiased. These feelings or sentiments are of high incentive to organizations, establishments, and different associations to enhance their administrations and items in view of clients' input.

Keywords: NLP, Deep learning, CNN, Sentiment analysis, Twitter data

I. INTRODUCTION

Deep Learning is a subset of machine learning which tries to mimic the natural way the humans use to gain knowledge of certain type. The algorithms used in deep learning are inspired by the functioning of the human brain. Deep learning, when compared to traditional machine learning, provides more accuracy as well as it is domain independent. In traditional machine learning model, the programmer uses labeled data and teaches the model to perform feature extraction. Whereas in the deep learning model when fed with unlabeled data it automatically features extracts as over a period of time the model's accuracy improves based on the data fed to it.

II. GENERAL STEPS FOLLOWED IN NLP PROJECTS

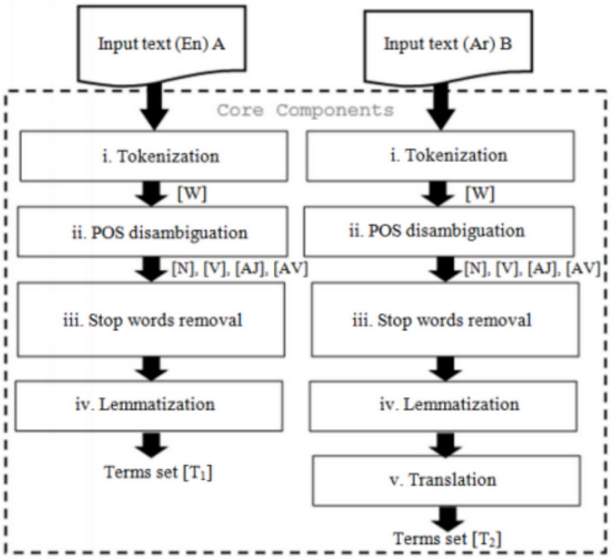


The above figure shows the general steps followed in NLP project. Where initially data is fetched from websites, blogs, social media. Which is then pre-processed and fed into the model to obtain the predictions. The accuracy and predictions can vary based on many factors such as pre-processing techniques, the algorithm used, data been fed into the model as well as the model training procedures.

II. RELATED SURVEY

Past research on pre-processing of data

Text pre-processing is a basic piece of NLP framework since the characters, words, and sentences recognized at this stage are key units on which classification and prediction happens. Pre-processing is performed initially as it helps us in eliminating noise from the dataset, later helps us in identifying the root word and consequently diminishing the size of the text data and enhancing the execution of the system. Any data set used for NLP is first tokenized i.e. the stream of content is separated to words or expressions or symbols. There are various challenges faced during the tokenization phase such as some languages like French, English are space-delimited, whereas Chinese and Thai are referred to as unsegmented as words do not have clear boundaries. Tokenization of such languages requires additional lexical and morphological details.



In the above figure shows the pre-processing of English and Arabic data. After POS disambiguation phase the data gets split into Nouns, Verbs, Adjectives and Adverbs. In Arabic language there is a requirement of translation to English for ease in training of the model.

After tokenization is performed, stop words are removed, followed by lemmatization and stemming is performed. Stop words are the words which are very frequently used such as - ‘and’, ‘are’, ‘this’ etc. They are not useful in classification of documents. So, they must be removed [3]. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma

Past research on NLP using supervised approach

The vast majority of the past research in this area concentrated on utilizing one of the three classifiers like SVM, Naïve Bayes and Maximum Entropy [1]. Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. In spite of its effortlessness and the way that its conditional independence assumption clearly does not hold in certifiable circumstances, Naive Bayes-based content classification still will, in general, perform shockingly well [4]. Maximum entropy characterization (MaxEnt, or ME) is an elective method which has demonstrated compelling results in various NLP application, sometimes it even outperforms Naïve Bayes but not always [5]. Support vector machines (SVMs) have been appeared to be exceptionally viable at traditional text categorization, for the most part outflanking Naive Bayes[6].

Even though these algorithms are most popular in Machine learning, it does not create great exactness in the wistful examination when contrasted with its application on traditional topic-based categorization [5]. The principle explanation behind such critical machine learning algorithms to fall flat is that in conventional content categorization strategies considered just terms i.e., unigram features and not the relation between them. One more approach is to represent content/text as Bag of Words (BOW) using unigrams. In any case, in Unigram-based portrayal, there won't be any thought for connection between unigrams, or in other terms, meaning or context of the sentences is not considered during the analysis of sentiments [7].

One answer for the above issue is the utilization of word sets. In the strategy, nouns are extracted with their modifiers. Phrases are represented by an abstraction called Head/Modifier pairs. As opposed to simply tossing expressions and catchphrases together, they begin with unadulterated HM sets and slowly add more keywords to represent the document

DOI 10.17148/IJIREEICE.2019.7206

[7]. Indeed, even in this strategy context or meaning of sentences are not considered as well as some words which provide polarity to content does not necessarily need to be in pairs. In machine learning, there is a number of classifier models available. These have colossal potential and can likewise give better outcomes. Distinctive machine learning classifier has its very own arrangement of design parameters, which are required to be tuned previously before the model gets prepared/trained. These parameters are known as hyperparameters. In the event that these hyperparameters are tuned appropriately model can give state of the art result for a problem else, it will do only exercise in futility. One such classifier used on movie reviews producing an accuracy of 87.85% is Random forest. One issue with this approach is that configuring the hyperparameters is tedious and complex.[1]

Past research on NLP using semi-supervised approach.

The extremity of a large portion of the sentences can be recognized by catchphrases alone. But the challenging part of NLP is when sentiments are expressed with inconspicuous semantic systems, for example, the utilization of mockery and very area particular relevant. For precedent, despite the fact that the sentence "The thief tries to protect his excellent reputation" contains the word "excellent", it reveals to us nothing about the creator's feeling and in actuality could be very much implanted in a negative audit [2].

This persuades the errand of learning vigorous sentiment models from minimal supervision. Recently The ongoing abundance of unlabeled information and less labeled information has drawn more and more consideration towards the models which can be prepared in a semi-supervised or unsupervised way. This is the place neural systems and deep learning architecture becomes possibly the most important factor. Active learning is another way that can limit the amount of required stamped data while getting engaged result. Generally, the preparation set is picked self-assertively. In any case, active learning picks the preparation information actively, which decrease the necessities of named data. As of late, active learning had been connected in sentiment classification. ADN architecture utilizes a new deep architecture for classification, and an exponential loss function aiming to maximize the separability of the classifier. ADN algorithm also identifies a small number manually labeled reviews by an active learner, and then trains the ADN classifier with the distinguished named information and the majority of the unlabeled information. In this way required marked information for preparing the model is lessened which is more efficient and sensible.[2]

In the work done by Shusen Zhou, Qingcai Chen and Xiaolong Wang they used multiple datasets which contained reviews Movie (MOV), books (BOO), DVDs (DVD), electronics (ELE), and kitchen appliances (KIT). Each dataset includes 1,000 positive and 1,000 negative reviews. They achieved over 75% accuracy in three of these five datasets which shows the success of the semi-supervised approach of training. [2] Another deep learning, feedforward network which is used in sentence classification is Convolutional Neural Networks (CNN) which is used by Yoon Kim. This avoids the disadvantages caused by active learning such as more time to prepare the model, domain-specific of the trained model as well as less efficiency. In this work we observe that a basic CNN with little hyperparameter tuning and static vectors accomplish magnificent outcomes on multiple benchmarks. [9].

CNN utilizes layers with convolving filters that are applied to detect local features. This was originally invented for computer vision. The approach done in this paper is - First the model is trained using a simple CNN with one layer of convolution on top of words vectors obtained from an unsupervised language model. These vectors were trained by Mikolov et al (2013) on 100B words of google news. After little tuning of hyper-parameters model achieves excellent results.[9]

Several variants of the model were made:

- CNN-rand: A model where all words are randomly initialized
- CNN-static: A model with pre-trained vectors.
- CNN-non-static: Same as above but the pre-trained vectors are fine-tuned for each task.
- CNN-multichannel: A model with two sets of word vectors. [9]

Datasets used for testing are movie reviews with one sentence per review, subjectivity dataset where the task is to classify a sentence as subjective or objective, TREC question dataset which involves classifying a question into 6 question types and customer reviews (CR) for various products [9]. With one layer of convolution performs exceptionally well and it adds to established evidence that unsupervised pre-training of the model gives better results in deep learning. [9]

The work is done on Sentiment Analysis In Hindi by Naman Bansal and Umair Z Ahmed [8] using Deep Belief Network to analyze sentiments in Hindi language of movie review dataset which they collected manually. They used DBN in the semi-supervised approach as the Hindi language is morphologically rich and is a free order language as compared to English as well as due to the scarcity of labeled data in Hindi for training the model.

DOI 10.17148/IJIREEICE.2019.7206

Deep Belief Networks are like neural networks yet they contrast in the quantity of covered up layers. Deep Belief Networks have numerous hidden layers stacked one over the other to catch the complex nonlinearity in the data. They basically unravel the fundamental values in the data. Deep Belief Networks are utilized to catch the fundamental nonlinearity which explains the variety of data. Deep Belief Networks have various layers so they can without much of a stretch capture the complex nonlinearity in the variety of data [8]. The main principle inspiration for utilizing the Deep Belief Networks with the unsupervised approach is to exploit unlabeled data which is available in abundance as well as to make the model less domain-specific. In the work done on Sentiment Analysis in Hindi [8] semi-supervised approach of training the DBN model, it was divided into two stages:

- In the first stage, pre-training of the model is performed in greedy layer-wise unsupervised learning manner with lots of unlabeled data and few labeled data.
- In the second stage, fine-tuning of the model is done using the gradient descent method.

The semi-supervised approach of training a deep learning model is possibly the most efficient manner of training. It also complements the current scenario where we have less labeled data and more of unlabeled data.

Past research done on NLP using Twitter data

To make the model less domain-specific most of the researchers uses live tweets from Twitter API. Twitter is a popular microblogging service where users create status messages (called “tweets”). These tweets sometimes express opinions about different topics [4]. Tweets format is completely different and short which gives another level of challenge where we need to handle slangs, abbreviations etc. [10]. Collection of these tweets can be done using Twitter API and secret keys are given by the API when we register our application.

The work was done by Pranav Waykar on Sentiment analysis in Twitter using Natural Language Processing (NLP) and classification algorithm as well as by Bhumika Gupta on Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python provide us with in-depth knowledge on the preprocessing that needs to be done on the data fetched from Twitter API. [10]

As all other pre-processing of data for NLP application, all the tweets are converted to lowercase followed by removal of stop-words. After which Twitter features and URL are removed. As Twitter limits the number of characters to be used most of the tweets contains URL to news articles or blog posts. It can also contain usernames of the accounts which are tagged along with the tweet.

These URL and usernames are non-beneficial to our sentimental analysis due to which we remove it out in the pre-processing stage [4]. After URL and links are removed, expansion of slangs and abbreviations used in tweets followed by stemming and removal of digits and special characters which does not provide any value to sentimental analysis [10].

Some tweets’ polarity may depend on the perspective you are interpreting the tweet from. For example, in the tweet “Federer beats Nadal :)”, the sentiment is positive for Federer and negative for Nadal. In this case, semantics may help. Using a semantic role labeler may indicate which noun is mainly associated with the verb and the classification would take place accordingly. This may allow “Nadal beats Federer :)” to be classified differently from “Federer beats Nadal :)”. [4]

IV. CONCLUSION

From the knowledge we gained by referring multiple project reports and survey papers belonging to the domain of Natural language processing, we conclude that semi supervised approach with deep learning architecture model is best suited for training the model when there is massive availability of unlabeled data with few labeled data. It also suits for models which are meant not to be domain specific.

REFERENCES

[1]. Hitesh Parmar, Sanjay Bhanderi and Glory Shah ‘Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters’ Conference Paper · July 2014

[2]. Shusen Zhou, Qingcai Chen, and Xiaolong Wang. Active deep networks for semi-supervised sentiment classification. Coling, 2010.

[3]. Dr.S.Kannan and Vairaprakash Gurusamy ‘Preprocessing Techniques for Text Mining’, October 2014

[4]. Pranav Waykar, Kailash Wadhwani, Pooja More ‘Sentiment analysis in twitter using Natural Language Processing (NLP) and classification Algorithm' International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016

[5]. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Conference on Empirical Methods in Natural Language Processing.

[6]. Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, Ms. Kalpana Dange ‘Sentiment Analysis Using Support Vector Machine’ International Journal of Innovative Research in Computer and Communication Engineering, January 2014

DOI 10.17148/IJIREEICE.2019.7206

[7]. Soumya George K , Shibily Joseph 'Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature' IOSR Journal of Computer Engineering (IOSR-JCE), Jan. 2014
[8]. Naman Bansal and Umair Z Ahmed 'Sentiment Analysis In Hindi' IITK, 2013
[9]. Yoon Kim 'Convolutional Neural Networks for Sentence Classification' New York University, 2014
[10]. Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani 'Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python' International Journal of Computer Applications, 2017.

DOI 10.17148/IJIREEICE.2019.7206