Current Twitter Trends Sentiment Analysis using Convolution Neural Network

Dr. M.Pauline, Annu S Ninan, Bilwa Gutthi, Deepika p and Kanchan kumari
Department of Computer Science Engineering
Gopalan College of engineering
Hoodi, Bangalore 560048, INDIA
{annuninan12345, bilwa.gutthi, deepikapk0918 & kanchan789456}@gmail.com

***Abstract* - Twitter is a micro-blogging system that allows you to send and receive short posts called tweets. Twitter has become the best indicator of the wider pulse of the world and what's happening within it. Analyzing the nature of these tweets can be helpful in fields like business, sociology, economics, and physiological studies. Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. When many Twitter users tweet about the same topics at the same time, they "trend" or become "trending topics". Twitter then lists the current top 30 topics on the trending page. Topics break into the Trends list when the volume of Tweets about that topic at a given moment dramatically increases. Performing sentiment analysis on these topics helps to analyze whether the reaction to topic is positive or negative. This is done using a convolution neural network. The network is trained with a dataset which has tweets already labeled as positive and negative. Then the top tweets for current trending topics are gathered and passed through the neural network to determine whether the trend has more positive or negative reactions.**

Index Terms – Sentiment analysis, Twitter trends, Deep Neural Network.

I. Introduction

Twitter is an American microblogging and social networking service on which users post and interacts with messages known as "tweets".

A word, phrase or topic mentioned in tweets at greater rate than others is known as a trending topic or twitter trend. They become popular by mutual efforts of many users or certain event that makes people to tweet about it. These trending topics help Twitter and their users to understand what is happening in the world and what people's opinions are about them. As users interact via social media spaces, like Twitter, they form connections that emerge into complex social network structures.

Twitter aggregates these trending topics and displays them on the trending page. There are two platforms on which the trends are worldwide and locale-based platforms. The world wide trends show the hash-tag or topic that is trending throughout the world and locale trends show the trends that are popular in that locality.

Sentimental Analysis is the process of determining whether tweet can be classified as positive, negative or neutral by gathering all the top trending on twitter both world wide and locale. About 30 topics trends are chosen by twitter as the top trending topics. The Trends list also displays the hottest emerging topics, not just what's most popular.

Convolution Neural Networks (ConvNets or CNNs) are a category of Neural Networks that have proven very effective in areas such as image recognition and classification. In sentiment analysis, the CNN is trained using multiple datasets that are tweets which are classified as positive or negative.

The system developed will take 30 current trends, from which 100 top tweets from each of the trends are taken from twitter. These 100 tweets are run through the convolution neural network. The classification of each tweet is obtained. The classification is aggregated for each topic. Thus, the polarity of the topic is obtained.

II.LITERATURE SURVEY

Sentiment analysis over Twitter offer organisations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. This is a novel approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. "Apple product") as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment. We apply this approach to predict sentiment for three different Twitter datasets. Our results show an average increase of F harmonic accuracy score for identifying both negative and positive sentiment of around 6.5% and 4.8% over the baselines of unigrams and part-of-speech features respectively. [2]

Sentiment analysis is a growing area of research with significant applications in both industry and academia. Most of the proposed solutions are centered around supervised, machine learning approaches and review-oriented datasets. In this article, we focus on the more common informal textual communication on the Web, such as online discussions, tweets and social network comments and propose an intuitive, less domain-specific, unsupervised, lexicon-based approach that estimates the level of emotional intensity contained in text in order to make a prediction. Our approach can be applied to, and is tested in, two different but complementary contexts: subjectivity detection and polarity classification. Extensive experiments were carried on three real-world datasets, extracted from online social Web sites and annotated by human evaluators, against state-of-the-art supervised approaches. The results demonstrate that the proposed algorithm, even though unsupervised, outperforms machine learning solutions in the majority of cases, overall presenting a very robust and reliable solution for sentiment analysis of informal communication on the Web.[3]

The use of ensemble learning, deep learning, and effective document representation methods is currently some of the most common trends to improve the overall accuracy of a text classification/categorization system. Ensemble learning is an approach to raise the overall accuracy of a classification system by utilizing multiple classifiers. Deep learning-based methods provide better results in many applications when compared with the other conventional machine learning algorithms. Word embeddings enable representation of words learned from a corpus as vectors that provide a mapping of words with similar meaning to have similar representation. In this study, we use different document representations with the benefit of word embeddings and an ensemble of base classifiers for text classification. The ensemble of base classifiers includes traditional machine learning algorithms such as naïve Bayes, support vector machine, and random forest and a deep learning-based conventional network classifier. We analysed the classification accuracy of different document representations by employing an ensemble of classifiers on eight different datasets. Experimental results demonstrate that the usage of heterogeneous ensembles together with deep learning methods and word embeddings enhances the classification performance of texts.[4]

Building high accuracy text classifiers is an important task in biomedicine given the wealth of information hidden in unstructured narratives such as research articles and clinical documents. Due to large feature spaces, traditionally, discriminative approaches such as logistic regression and support vector machines with n-gram and semantic features (e.g., named entities) have been used for text classification where additional performance gains are typically made through feature selection and ensemble approaches. A more direct approach using convolutional neural networks (CNNs) outperforms several traditional approaches in biomedical text classification with the specific use-case of assigning medical subject headings (or MeSH terms) to biomedical articles. Trained annotators at the national library of medicine (NLM) assign on an average 13 codes to each biomedical article, thus semantically indexing scientific literature to support NLM's PubMed search system. Recent evidence suggests that effective automated efforts for MeSH term assignment start with binary classifiers for each term. CNNs are used to build binary text classifiers and achieve an absolute improvement of over 3% in macro F-score over a set of selected hard-to-classify MeSH terms when compared with the best prior results on a public dataset. Additional experiments on 50 high frequency terms in the dataset also show improvements with CNNs.Results indicate the strong potential of CNNs in biomedical text classification tasks.[5]

Twitter sentiment analysis offers organizations an ability to monitor public feeling towards the products and events related to them in real time. Most existing researches for Twitter sentiment analysis are focused on the extraction of sentiment feature of lexical and syntactic feature that are expressed explicitly through words, emoticons, exclamation marks etc, although sentiment implicitly expressed via latent contextual semantic relations, dependencies among words in tweets are ignored. In this paper, we introduce distributed representation of sentence that can capture co-occurrence statistics and contextual semantic relations of words in tweets, and represent a tweet via a fixed size feature vector. We used the feature vector as sentence semantic feature for the tweet. We combined semantic feature, prior polarity score feature and n-grams feature as sentiment feature set of tweets, and incorporated the feature set into Support Vector Machines(SVM) model training and predicting sentiment classification label. We used six Twitter datasets in our evaluation and compared the performance against n-grams model baseline. Results show the superior performance of our method in accuracy sentiment classification.[6]

Sentiment analysis is concerned with the automatic extraction of sentiment-related information from text. Although most sentiment analysis addresses commercial tasks, such as extracting opinions from product reviews, there is increasing interest in the affective dimension of the social web, and Twitter in particular. Most sentiment analysis algorithms are not ideally suited to this task because they exploit indirect indicators of sentiment that can reflect genre or topic instead. Hence, such algorithms used to process social web texts can identify spurious sentiment patterns caused by topics rather than affective phenomena. This article assesses an improved version of the algorithm SentiStrength for sentiment strength detection across the social web that primarily uses direct indications of sentiment. The results from six diverse social web data sets (MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums) indicate that SentiStrength 2 is successful in the sense of performing better than a baseline approach for all data sets in both supervised and unsupervised cases. SentiStrength is not always better than machine-learning approaches that exploit indirect indicators of sentiment, however, and is particularly weaker for positive sentiment in news-related discussions. Overall, the results suggest that, even unsupervised, SentiStrength is robust enough to be applied to a wide variety of different social web contexts.[7]

III. SYSTEM OVERVIEW

This section explains the steps which are carried out in order to get the sentiment analysis of current twitter trends

A. Setting up Twitter API

The first step in setting up a connection to twitters website in order to obtain information required such as the current trends and tweets to perform sentiment classification.

Twitter provides as API which can be accessed by the python library "tweepy". This library provides simple automation and creating twitter bots.

The twitter API can be used through a twitter developer account. Once set up, authorization key is provided by which the system can access the functions of the API.

B. Getting current trends

Once the API is set up, we can use its function to obtain the current trends. Trends of any place can be obtained by using the locale number of the locality from which we want to find the trends. In this paper we have used the trends from the United Kingdom, an English speaking locality. As the neural network is developed using words from the English language it would be ideal for most of the trend words to be in English. Using an English speaking locality such as the United Kingdom gives a higher probability of the trends being in English.

Once the current trends have been obtained they are stored in a list.

C. Getting tweets related to the current trend

Using the list of trends we use the twitter API to search each trend keyword from the list. The top hundred tweets are collected and placed in a dictionary data-structure.

D. Performing sentiment analysis

From the gather tweets sentiment analysis is performed. To perform sentiment analysis the tweets are passed through a convolution neural network and the classification is obtained. The working of the convolution neural network is explained in Section [III]. The classification is obtained for the tweets are aggregated and the positive percentage and the negative percentage of each trend is output.

III. Developing the Convolution Neural Network

The first layers embeds words into low-dimensional vectors. The next layer performs convolutions over the embedded word vectors using multiple filter sizes. The system does not use pre-trained word2vec vectors for word embeddings. Instead, the system learns embeddings from scratch. The system does not enforce L2 norm constraints on the weight vectors.

A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification found that the constraints had little effect on the end result. The paper experiments with two input data channels – static and non-static word vectors. We use only one channel. The model is implemented as follows, to allow various hyperparameter configurations the system puts our code into a TextCNN class, generating the model graph. A CNN for text classification. Uses an embedding layer, followed by a convolutional, max-pooling and soft-max layer. The system considers, the length of our sentences, number of classes in the output layer, two in our case (positive and negative), the size of our vocabulary, the dimensionality of our embeddings, the number of words we want our convolutional filters to cover, the number of filters per filter size.

The system inputs placeholders and starts by defining the input data that we pass to our network. The probability of keeping a neuron in the dropout layer is also an input to the network because we enable dropout only during training. We disable it when evaluating the model. The first layer the system defines is the embedding layer, which maps vocabulary word indices into low-dimensional vector representations. It's essentially a lookup table that we learn from data. Then the system builds our convolutional layers followed by max-pooling. We use filters of different sizes. Because each convolution produces tensors of different shapes, we need to iterate through them, create a layer for each of them, and then merge the results into one big feature vector. Dropout is the perhaps most popular method to regularize convolutional neural networks. The idea behind dropout is simple. A dropout layer stochastically "disables" a fraction of its neurons. This prevents neurons from co-adapting and forces them to learn individually useful features.

The system set this to something like 0.5 during training, and to 1 (disable dropout) during evaluation. Using the feature vector from max-pooling (with dropout applied) we can generate predictions by doing a matrix multiplication and picking the class with the highest score. We could also apply a soft-max function to convert raw scores into normalized probabilities, but that wouldn't change our final predictions. Using our scores, we can define the loss function. The loss is a measurement of the error our network makes, and our goal is to minimize it. The standard loss function for categorization problems is the cross-entropy loss. We also define an expression for the accuracy, which is a useful quantity to keep track of during training and testing.

[6] Sentimental analysis can be used for automatic extraction of sentimental-related information from text . Even though sentimental analysis is used for commercial tasks like extracting opinions from product reviews, there is increase in popularity in effective dimensions of social web that is twitter in particular. Not all the algorithm used in sentimental analysis can be used to do such tasks as the use indicators of sentiments indirectly that may affect the genre or topic instead. So, these algorithms are used to determine spurious sentiment patterns that are caused by topics from social web texts.

[7] Polarity classification in twitter uses supervised approach where the only difference between them is the features selected and methods used for weighing them. The results obtained show that disambiguation and expansion are good strategies for improving overall performance.

[8]

References

[1] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," in *IEEE Access*, vol. 6, pp. 23253-23260, 2018.

[2] H. Saif Y. He H. Alani "Semantic sentiment analysis of twitter" Proc. Semantic Web-ISWC pp. 508-524 2012 2012.

[3] G. Paltoglou M. Thelwall "Twitter MySpace Digg: Unsupervised sentiment analysis in social media" ACM Trans. Intell. Syst. Technol. vol. 3 no. 4 pp. 1-19 2012.

[4] N. Kalchbrenner E. Grefenstette P. Blunsom "A convolutional neural network for modelling sentences" Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics vol. 1 pp. 655-666 Jun. 2014.

[5] Y. Kim "Convolutional neural networks for sentence classification" Conf. Empirical Methods Natural Lang. Process. (EMNLP) pp. 1746-1751 Oct. 2014.

[6] Z. Jianqiang, C. Xueliang, "Combining semantic and prior polarity for boosting twitter sentiment analysis", Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity), pp. 832-837, Dec. 2015.

[7] . M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection for the social Web", J. Amer. Soc. Inf. Sci. Technol., vol. 63, pp. 163-173, 2012.

[8]