ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 4, April 2015

Harnessing Twitter Big Data for Automatic Emotion Identification

R.Rajeswari

PG Scholar, Computer Science and Engineering, PSG College of Technology, Coimbatore, India

Abstract: Sentiment analysis is used to identify and extract subjective information (verb, noun, adjectives) in text sentences. The dataset used for emotion identification is collected from twitter. User posting contents in twitter is increasing every day. Emotion identification is one of the concept for understanding users behaviour, action and is important to all phases of our life. The emotions are automatically labelled according to the emotion hashtag. Each information recovered from twitter has separate hashtag. Naive Bayes (NB) machine learning algorithm is used for emotion identification. To find valuable features for emotion identification like n-grams, LIWC Dictionary, part-of-speech (POS). Finally the extracted Features is fed into a Naive Bayes (NB) classifier to attain classification accuracy on emotion identification in Twitter dataset.
Naïve Bayes classification is a simple probabilistic model that works well on text classification. By using Twitter content, it classifies the different emotions like joy, sadness, anger, love, fear, thankfulness, and surprise for the twitter content.

Keywords: Naïve Bayes(NB) classifier,N-Gram, LIWC Dictionary, Part-of-speech.

I. INTRODUCTION

Emotions are important to all features of our life. It helps our decision-making, which involves social relationships, profiles of daily behaviour and still outlives our memories.

Due to the fast development of emotion-rich textual content, such as micro blog posts, blog posts, and environment discussions, there is a great need and chance to expand usual tools for identifying and exploring people's emotions expressed in the text. Very challenging factor here is identifying the emotions in text. The emotions can be understood and activated by exact events or situations.

The text relates to an event or situation that has the emotion can be lacking clear emotion-behaviour words and we cannot identify fear emotion if there is no clear reference to words such as "shock" and "fear". Identification of different emotion is based on the source of the keywords on earlier. But gathering division between different emotions is purely on the source of keywords but the current emotion identification research mainly on learned training data.

The understanding of the data by human experts is very effort-challenging and time consuming. Moreover, in the difference with other similar tasks such as article or topic detection how a human understands an emotion in the text has a tendency to be subjective sentences. Twitter is the popular micro blogging service, gives more than 340 million tweets per day on a large variety of issues, and a major part of it is about "what is happening" in our daily lives expressed using emotion hash tags.

In the following example, "I won first price in basket ball" is the tweet sentence. This depicts #happy emotion. Thus the user "understands" the tweet with the hash tag #happy to be express tenseness emotion. Can this Twitter 'big data' be controlled to undertake the emotion identification difficulties? The following three problems are addressed. First problem is automatically creating a big emotion dataset with high quality tags from Twitter by controlling the emotion hash tags. Second problem is understand what features can successfully improve the performance of managed machine learning algorithms. Third problem is does big (millions as a substitute of thousands) training data improve emotion identification truthfulness and how much performance gain can be reached by rising the size of training data.

To improve the excellence of collected tweets, a set of methods were developed to maintain the applicable tweets, which contain the emotion hash tags that the correctly understand the expressed emotions.

To find valuable features for emotion identification, a large variety of the features including n-grams, emotion Dictionary, part-of-speech (POS), n-gram positions, etc were searched. Machine learning algorithm namely Naive Bayes (NB) were used.

II. RELATED WORK

The text-based emotion identification problem in the area of children's fairy tales, with child-directed communicative text-to-speech separation as goal was introduced [2]. It explores the text-based emotion identification problem using the supervised machine learning with the SNoW learning architecture. The goal is to classify the emotional similarity of sentences in the storyline area of children's fairy tales, for resultant usage in proper expressive representation of text-to-speech separation.

Initially their experimented on a preface data set of 22 fairy tales and showed hopeful results over a naive baseline and BOW approach for classification of

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 4, April 2015

emotional opposed to non-emotional contents, with some dependence on restriction modification.

Emotions are poorly understood, and it is particularly unclear which features may be important for their identification from text. Thus, they experimented with different feature designs. Starting with all features, again using 10-fold cross-validation for the separated modification-evaluation condition, one additional feature group was removed pending none continued.

The benefit is first, what emotion or emotions most properly describe a certain text passage, and second, given a text passage and a particular emotional mark-up, how to provide the prosodic form in order to transmit the emotional content. The text-based emotion identification task (TEP) addresses the first of these two problems.

The drawback of "Emotions from text: machine learning for text-based emotion prediction" is removing any group corrupted performance because features interact and there is no true independence. It was observed that features offerings were sensitive to parameter modification.

An English POS tagger that is designed for Twitter data was introduced [3]. They developed a POS tag set for Twitter, They physically tagged 1,827 tweets, developed features for Twitter POS tagging and performed experiments to calculate them.

Twitter includes many alternate spellings of words; the authors use the metaphone algorithm to create a common phonetic normalization of words to simpler keys. Metaphone consists of 19 rules that rewrite consonants and delete vowels. The feature is WORTH: Twitter orthography, NAMES: Frequently-capitalized tokens, TAGDICT: Traditional tag dictionary, DISTSIM:Distributional similarity. METAPH:Phonetic normalization.

The benefit is their tagger with the full feature set achieves a relative error reduction of 25% compared to the Stanford tagger.

The drawback of "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments" is recall of proper nouns is only 71%.

The Waikato Environment for Knowledge Analysis (WEKA) came about through the professed need for a joined workbench that would allow researchers easy access to state-of the-art techniques in machine learning was introduced [4].

Learning algorithms were available in various languages, for use on different platforms, and operated on a variety of data formats. The task of collecting together learning methods for a relative study on a collection of data sets was discouraging. It was visualized that WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be worried with supporting infrastructure for data management and design estimation.

The most important new features in WEKA 3.6 are centers are Learning Schemes, Pre-processing filters, and User interfaces, Extensibility.

The benefit is it was decided that advantages such as "Write Once, Run Anywhere"

The drawback of " WEKA Data Mining Software: An Update" is considering running time, there is no longer a major disadvantage compared to programs written in C, a commonly-heard argument against Java for data-thorough processing tasks, due to the complexity of just-in-time compilers in modern Java virtual machines.

A linguistic resource for a lexical representation of affective knowledge was introduced [5]. This resource was developed starting from WORDNET, through the selection and classification of the synsets representing affective concepts.

The methodology at first attempts to build a lexical structure for affective terms concerned studying which conditions are really representing emotions, and what classification criteria to consider. In particular, lexical semantic approaches are founded on the principle that it is possible to gather emotion properties from the emotion words .This approach consists of three main steps. First, emotion words are collected from dictionaries or from literary and newspaper texts. Then, a fixed number of semantic environments are fixed. For example pure emotion words, personality characteristic words etc. Finally, from each word a set of affective measurements is extracted, using techniques such as factorial analysis or multidimensional scaling.

The benefit for word net-affect is useful in all applications in which it is compulsory to have an affective interaction.

The drawback of" WorldNet-Affect: an Affective Extension of Word Net" is Number of affective synsets is less.

A data-oriented method for gathering the emotion of a speaker conversing with a dialog system from the semantic content of a declaration was introduced [6]. We first fully automatically obtain a large collection of emotion-provoking event instances from the Web.

The methodology is that they first automatically collected a large collection, as many as 1.3M, of emotion-provoking event occurrences from the Web. Then the emotion classification task is decomposed into two sub-steps: emotion polarity classification and emotion classification. In emotion polarity classification, we used the EP-corpus as training data.

The benefits of this approach are that it is generally known that performing fine-grained classification after common classification often provides good results particularly when the number of the classes is large.

The drawback of "Emotion Classification Using Massive Examples Extracted from the Web" is that the identification of neutral sentences is very less.

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 4, April 2015

III. SYSTEM DESIGN

The overall System architecture diagram is shown in Fig 1.

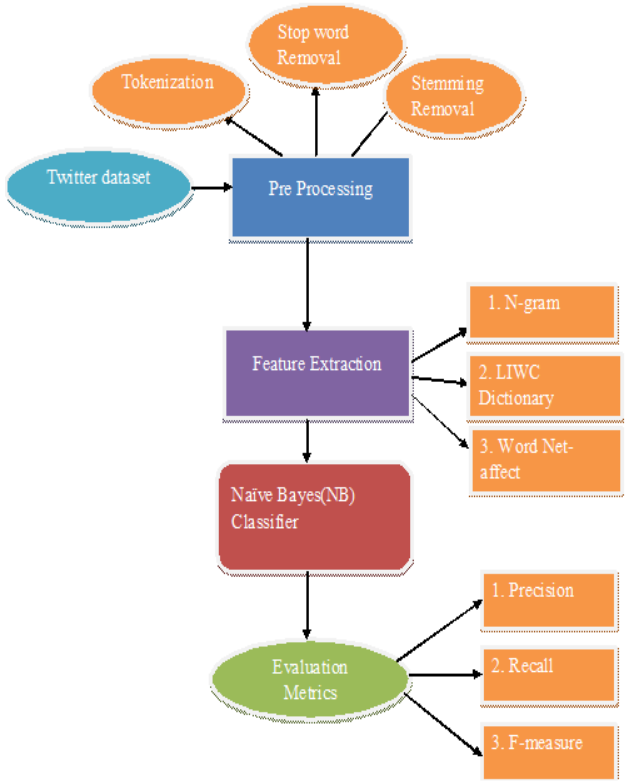

Fig.1. Overall system architecture

Twitter Dataset contain four emotions like anger, happy ,fear ,sad in 546 Text sentences, First pre-process was performed for given twitter dataset. Following are the pre-processing also Tokenization, stop word and Stemming was performed for the given twitter dataset. Feature extraction was performed using n-gram model, Part of Speech and LIWC Dictionary. Finally Naïve Bayes(NB) Classifier was used to classify the accuracy for given twitter dataset which is based on the following measures: F-measure, precision, recall.

IV. CLASSIFIER

A. Naive Bayes(NB) Classifier

A statistical classifier called Naïve Bayesian classifier is based on the Bayes Theorem. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved.

Naïve Bayes classification was given by the following formula in (1).

$$p(c_i|X)=\frac{p(X|c_i)p(c_i)}{p(X)} \quad \ldots.(1)$$

Where,

D is a training set of n records, and record i is represented as (X, c_i) where X= (x_1, x_2, …, x_m). If there are k classes c_1, c_2, …, c_k.

Classification is to derive the maximum posteriori, i.e., the maximal $p(c_i|X)$.

This can be derived from Bayes' theorem(2).Since p(X) is constant for all classes, only needs $p(X|c_i)p(c_i)$to be Maximized

After finishing preproccesing and feature extraction steps in twitter dataset. Naïve Bayes classifier is used to classify accuracy for this dataset using following evaluation metrics.

V. IMPLEMENTATION / PERFORMANCE ANALYSIS

There are three modules. First module is preprocessing, the second module is feature extraction. Feature extraction module implements these modules. N-gram (unigram, Bigram) model and LIWC Dictionary, Parts-of-speech.

A.Preprocessing

Following are the steps performed in data preprocessing.

First it checks letters or punctuations that are repeated more than twice were replaced with the same two letters punctuations or punctuations (e.g., cooool -> cool, !!!!! -> !!), frequently used informal expressions were normalized and hash symbols were retained.

Then it performs Tokenization (splitting the words), stop word removal (e.g.Is, That, The, We, etc.) and Stemming (e.g. Going, Goes -> Go.) was performed for the given twitter dataset.

B. Feature Extraction

It helps to visualize high dimensional data. It prevents the problem of over-fitting. It reduces noisy, redundant data and memory needed to store data.

It start with features that are known to be effective for text classification especially sentiment analysis. Sentiment and emotion are subjective comparative study of the useful features for identifying them may provide better insights for emotion identification.

(i) N-gram is combination of unigram (n=1), bigram(n=2) are used for emotion analysis. Emotions, punctuation, parts of speech are included in this model and are used for tokenization. Stop word removal is used in this model.

Unigram is an n-gram consisting of a single item for a sequence.

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 4, April 2015

In the table I Following Result depicts the performance of different emotion after performing unigram:

TABLE I

UNIGRAM-EMOTION ACCURACY

Class	Precision	Recall	F-Measure
sad	0.981	0.986	0.983
anger	0.974	0.972	0.973
happy	0.968	0.991	0.979
fear	1	0.892	0.943

Bigram is every sequence of two adjacent elements in a string tokens.

In the table II Following Result depicts the performance of different emotion after performing bigram.

TABLE II

BIGRAM-EMOTION ACCURACY

Class	Precision	Recall	F-Measure
sad	0.98	0.987	0.983
anger	0.968	0.991	0.979
happy	0.991	0.988	0.989
fear	1	0.91	0.953

(ii) Linguistic Inquiry and Word Count

(LIWC) Dictionary is text analysis software which covers about 4,500 dictionary words and these words were stems from more than70 categories. Here it collected emotion words from the positive emotion category (408 words) and negative emotion category

(499 words) in LIWC2007 dictionary. For each tweet, here it counted the number of positive/negative words based on the set of collected emotion words, and used the percentage of words that are positive and that are negative as features.

In the table III following Result depicts the performance of different emotion after performing LIWC Dictionary:

TABLE III

LIWC DICTIONARY-EMOTION ACCURACY

Class	Precision	Recall	F-Measure
sad	1	0.982	0.991
anger	1	0.983	0.991
happy	1	0.988	0.994
fear	0.772	1	0.871

(iii) Part-of-Speech(POS) is a class of words based on the word's function, the way it works in a sentence.

It will identify noun, verb, pronoun, adjective etc., in the twitter content.

In the table IV following Result depicts the performance of different emotion after performing POS:

TABLE IV

POS -EMOTION ACCURACY

Class	Precision	Recall	F-Measure
Sad	1	0.981	0.99
Anger	1	1	1
Happy	0.981	1	0.99
Fear	1	0.98	0.99

C. Evaluation metrics

Accuracy is overall performance of individual classifier is measured by:

$$\text{Accuracy} = \frac{\text{No of correctly labeled tweets}}{\text{Total no of Tweets in the dataset}}$$

Precision is the measurement of correctness.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Positive})}$$

Recall is the measurement of completeness.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Negative})}$$

F-measure is the harmonic mean of precision and recall.

$$\text{F-measure} = \frac{(2*\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})}$$

Table V depicts different features accuracy by Naïve Bayes classifier.

TABLE V

ACCURACY OF VARIOUS FEATURES

Features	Accuracy(%) of Naïve Bayes (NB) Classifier
n-gram(n=1)	97.61
n-gram(n=2)	98.18
LIWC Dictionary	98.54
Part-of-Speech	99.31

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 4, April 2015

VI. CONCLUSION

Thus the twitter dataset was performed pre-process(Stop word removal tokenization ,n-gram) and then feature extraction(LIWC Dictionary, Part-of-speech) and check the accuracy using Naive Bayes (NB) machine learning algorithm classifier. In future Naïve bayes can be compared with various other machine learning classification technic to check the accuracy.

REFERENCES

[1] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan,and Amit P. Sheth "Harnessing Twitter 'Big Data' for Automatic Emotion Identification" in 2012 International Confernece on Social Computing ,pp. 587 – 592.

[2] C. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in Proceedings of HLT and EMNLP.ACL, 2005, pp. 579–586.

[3] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein,M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of -speech tagging for twitter: annotation, features, and experiments," in Proceedings of HLT:short papers, ser. HLT '11. Strou dsburg, PA, USA: ACL, 2011, pp. 42–47.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.

[5] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in Proceedings of LREC, vol. 4. Citeseer, 2004, pp. 1083–1086.

[6] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion classification using massive examples extracted from the web," in Proceedings of COLING. ACL, 2008, pp. 881–888.

[7] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text,"in Proceedings of the 2008 ACM symposium on Applied computing.ACM, 2008, pp. 1556–1560.

[8] S. Mohammad, "#emotional tweets," in Proceedings of the Sixth International Workshop on Semantic Evaluation. ACL, 7-8 June 2012, pp.246–255.

[9] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006,pp. 27–29.

[10] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in Proceedings of IJCNLP, 2008, pp. 296–302.