I.J.Modern Education and Computer Science, **2013, 8, 51-57**
Published Online October 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2013.08.06

Artificial Neural Network in Prognosticating Human Personality from Social Networks

Harish Kumar V
PESIT South Campus, Bangalore, India
harrysofter@gmail.com

Arti Arya
PESIT South Campus, Bangalore, India
artiarya@pes.edu

Divyalakshmi V
Wipro Technologies, Bangalore, India
divijay1@gmail.com

Nishanth H S
Monsanto Holdings Pvt Ltd
nishanthmadhu@gmail.com

Abstract — The analysis of text in the form of tweets, chat or posts can be an interesting as well as challenging area of research. In this paper, such an analysis provides information about the human behavior as positive, negative or neutral. For simplicity, tweets from social networking site, Twitter, are extracted for analyzing human personality. Various concepts from natural language processing, text mining and neural networks are used to establish the final outcome of the application. For analyzing text, Neural Networks are implemented which are so modeled that they predict the Human behavior as positive, negative or neutral based on extracted and preprocessed data. Using Neural Networks, the particular pattern is identified and weights are provided to words based on the extracted pattern.Neural networks have an added advantage of adaptive learning. This application can be immensely useful for politics, medical science, sports, matrimonial purposes etc.The results so obtained are quite promising.

Index Terms — Neural Networks, Social Network Text Analysis, Text Mining, Wordnet

I. Introduction

There is a vast pool of unstructured text data available on Internet that is increasing exponentially day by day. Some of the most popular social networking sites are the great sources of such unstructured text data. This paper presents a useful way of analyzing such data for many decision making purposes. Suppose, a company wants to hire some professionals and at the same time they want to know about the basic personality traits of the people under consideration. In such a situation, the application proposed in this paper can help revealing whether the person is optimistic (positive), pessimistic (negative) or neutral. Such information can help the hiring company to decide whether the person would be an asset for the organization or not. Similarly, when parents look for a match for their son or daughter, they can use this application to get an insight into the basic attitude of the person. Here, an assumption is made that the person to be analyzed must have an account on Twitter (the social networking site considered for the results). So, this way the proposed application can help taking very important decisions in crucial scenarios.

The text is extracted and preprocessed (cleaned) for further analysis. For analyzing text, the multilayered artificial neural networks are implemented. The concept of Neural Networks is inspired from the human brain and nervous system. A neural network consists of information processing unit called a neuron. Neural networks are quite effective and competent with huge data. They are usually used to model intricate relationships between inputs and outputs to find underlying hidden patterns in textual data. Neural Networks have the ability to adapt to modified input so that the network produces result without the need to redesign the output criteria.

A Neural Network [1] is categorized based on many aspects. Some of the aspects are, its pattern of connection between the neurons called its architecture, its methods determining the weights on the connection

called its training and learning, algorithms and its Activation Function. Also, the design of these networks is parallel. Even if one neuron is not working, then also the classification results are obtained.

Also, Text Mining is a process of full or partial automation of exploring unknown hidden patterns in the text. It extracts relevant information from text, which is otherwise not very obvious [2]. In this paper, text has been put to preprocessing step to make it appropriate for further analysis. The various stop words are removed. The common words that occur quite frequently in text are unlikely to help mine the text. E.g., "the", "a", "an" are some of the stop words which are being removed. The words like "he, "she", "we", "them", "him" etc. are not removed as stop words because such words give the information about certain important features regarding the person under consideration.

The cleaned text is analyzed by using various similarity measures and semantic similarity. In text mining, Natural Language Processing also plays a vital role. Natural Language Processing (NLP) [3] is a computerized approach to analyze text based on various theories and technologies. Although it is a very active area of research and development, there is no single definition that would satisfy everyone.

The application proposed in this paper is best suited for the following fields but not limited to

- Criminal Sciences,
- Medical Science,
- Insurance Industry
- Sports
- Human Resource Department
- For Matrimonial purposes
- Politics etc.

The text is extracted from social networking site, Twitter and preprocessed by removing some stop words. Once the input text is ready, it is fed to the classifier and classifier classifies the person as positive, negative or neutral. Here classifier is developed using multilayered artificial neural networks. The motivation for implementing artificial neural networks for the purpose is that it adapts to the ever-changing input and accordingly provides the output.

The rest of the paper is organized as follows: Section II highlights the details of the literature available to the related field. Section III elaborates the idea of proposed framework in detail and the architecture of the system. In Section IV the outcomes of the proposed framework is discussed and snapshots are provided. Section V summarizes the whole idea along with the scope of future enhancements.

II. Related Work

Analyzing human personality through the text can be a real mystifying area of concern to a great mass. In this paper, implementation of artificial neural network with two hidden layers has done such kind of analysis.

Artificial neural network algorithms attempt to conceptualize the complexity of real biological system. The neural network's adaptability towards the training data makes it learn the behavior of training data thereby adjusting its weights iteratively [2].

Text Data Mining (TDM) or text mining [2] is a technique of revealing unknown patterns or nuggets in the large pool of text data. This text data is available in the form of unstructured text on webpages, links etc. online.

The neural networks do not accept categorical attributes directly but these attributes need to be mapped to numerical values [4]. The mapping so done in this way produces an ordering in the values that any neural network takes into account while processing the data. Also, the final results that are generated by a neural network are continuous values and are little difficult to interpret as categorical values. There are number of ways to interpret these values finally [4].

In [5], authors have used text mining and natural language processing concepts for analyzing text and thereby summarizing a large input text. The method proposed by the authors [5] has three main steps: input the text document, summary algorithm and output in the form of summary. For summary algorithm, the text is first preprocessed and then frequent terms are identified followed by filtering of the sentences for final summary.

In [6] authors have described about predicting gender of the persons who are chatting with each other. They have presented an idea of extracting the data from chatting sites and used the concepts of training the dataset and testing as steps towards developing classifier for predicting gender of the persons.

In [7], Hearnst has pointed the difference between TDM and Information retrieval. The author has also suggested the use of vast online text data to discover new patterns about the text itself. The author has also suggested that fully automated artificial intelligence based systems may not provide the expected results. So, a semi automatic system guided by human intervention may obtain better results. Fig 2 depicts the various steps involved in text data mining.

In [9], in order to evaluate chat conversation, the authors have collected the data from various chat resources. The extracted text is stored in text files and then preprocessed the data for text mining. Statistical

information regarding these chat conversations were also generated. Then, they have determined the characteristic of these chats and the topics of conversation. At last they have compared the results of conversation topic determination using Naïve Bayes, K- Nearest neighbor and support Vector Machines based classifiers.

In [10], authors have proposed a neural network classifier based on back propagation learning technique that does the cross validation for original neural network (BP algorithm). They have mainly focused on reducing the training time and increasing the accuracy of the classifier.

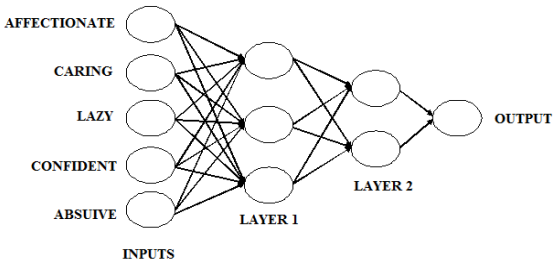

Fig. 1: Two Hidden Layers of Neural Networks

Artificial neural networks [11], [12] are composed of interconnecting artificial neurons and are used for solving artificial intelligence problems. A multi layered artificial neural network (with two hidden layers) is shown in Fig. 1. Here, five inputs are considered which are actually psychological parameters on the basis of which a human being is being analyzed.

The neural networks are widely being used for mining textual data also. Text Mining as shown in Fig 2 usually involves the process of structuring the input text, by automatically extracting information from a usually large amount of different unstructured textual resources and revealing hidden and unknown patterns in text data.

It is the automated or partially automated processing of text. It involves imposing structure upon text and extracting relevant information from text. Text mining methods are relevant to small document and as well as large document collection.

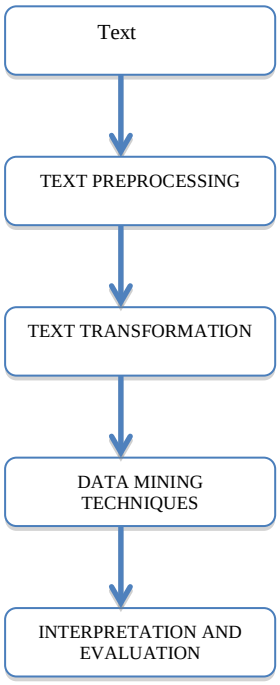

Fig. 2. Steps involved in text mining

In [13], authors have explored labeling human traits using various adjectives such as adorable, charming, abusive, pessimistic etc. They have used the concept of agglomerative clustering for arriving at the final result and have extracted tweets from twitter for labeling a person with several adjectives. In this paper, rather than using clustering for semantic similarity, we have implemented artificial neural networks to classify human personality as positive, negative or neutral. As artificial neural networks work efficiently even if they receive an input for which these networks are not trained. Also, while preparing the data for final input, it is taken care of, to whom the sentence is targeted at. Whether the user is making a general comment, he is talking about himself or talking about third person.

In this paper, the training of the Neural Network is carried out using Modified Madaline Algorithm [2], [11]. An extended algorithm for training multilayered fully connected feed-forward networks of ADALINE neurons is implemented. The algorithm is called MRII for MADALINE RULE [2].

III. Proposed Work

In the proposed framework, the text is extracted from the web and analyzed for determining human behavior using Text Mining techniques. A multilayered feed forward neural network does the classification of text. A neural network acts like a black box that processes the input and produces the output without giving any insight into “how processing is happening?” [8]. The system architecture is shown in Fig.3.

TABLE I Three Categories of Class Labels for Sentences

Affection-ate	Caring	Confidence	Absurd	Lazy
Honesty	Love	Positive	Idiot	Difficult
Trust-worthy	Dear	Sure	Mad	Tired
Fond	Share	Capable	Foolish	Tough
True	Darling	Done	Weak	Dull
Happy	Protect	Can	Insult	Impatient
Soft	Good	Strong	Goosey	Shy
Sweet	Provide	Promise	Foul	Sleepy
Warm	Help	Will	Dirty	Lifeless
Like	Friendly	Bold	Wrong	Drowsy
Concern	Thought	Secure	Arrogant	Drag
Loving	Attentive	Right	Evil	Bore
Supportive	Pity	Hope	Wild	Slow
Faith	Support	Trust	Vulgar	Sluggish

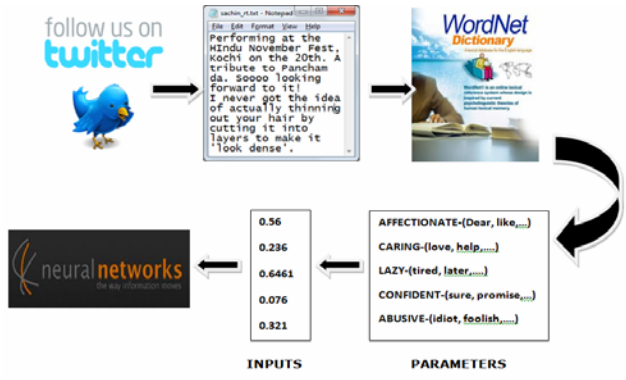

Fig. 3: The series of steps involved in analyzing a person.

As the outcome of the proposed framework, based on the text extracted, the human behavior can be predicted as positive, negative or neutral. Positive traits are affectionate, caring, confidence and so on. Negative traits if a person is lazy or abusive and neutral if the person cannot be classified into either one of the above based on the text extracted.

STEPS FOR ANALYZING A PERSON:

Step 1: Get the user name to analyze.
Step 2: Call Extract (username).
Step 3: Divide the complete extracted text into sentences.
Step 4: Divide the sentence in to words.
Step 5:Check to whom the tweet is targeted at. (Is it the person himself, others or in general) and set the flag to denote this.
Step 6: Find out if the sentence has any word that was meant to negate the meaning of the sentence like not, doesn't, don't, etc. and set the flag to denote this.
Step 7: Find the synonyms for the words in the sentence from Wordnet and compare it with the standard set of words categories earlier.
Step 8: A numerical value is associated with respect to each personality trait based on the number of words that match the standard set. And this numerical value is fed as input to neural network.
Step 9: Repeat steps 4 to 8 for all the sentences.
Step 10: Combine the results of all the individual sentences and generate a single value for each personality trait.
Step 11: Input those values to the Artificial Neural Network to get the output.

Algorithm: Extract (username)
Step1: Follow the person on twitter.
Step2: Get the recent tweets made by the person.
Step3: return

If the sentence has words such as "you, he, she, them, her, him, etc." then it most probably means that the speaker is targeting the sentence at another person or thing. If the word "we" appear in the sentence then the sentence was probably spoken in a general sense.

If the word "am" appears in the sentence then the speaker is targeting at himself. Finding this parameter help us understand the sentence better. So in order to identify to whom the sentence is targeted at, a simple rule has been incorporated. Three categories have been decided that are self, third person, general are shown in Table I.

The word "not", doesn't or won't normally negate the meaning intention of the sentence.

TABLE II Few Words Under Each Of The Basic Personality Parameter.

S.No	Category	Terms in the category
1.	Self	am
2.	Third Person	He, she, they, you, them, their, him, her
3.	General	We, us

Here one sentence is considered at a time for processing and a value is assigned to each term in the sentence. In order to compute this value, a synonym tree is constructed for every term. Then a term set is constructed with all the terms in the synonym tree for all the terms in the sentence. This set is then compared with standard term set for each personality trait as shown in table II. The match found is used to get the value corresponding to the personality.

Also, certain assumption are made to compute this value which are as follows:

- Only one term from each synonym tree is taken and the rest terms are all ignored.
- The term that appears first in the tree is considered.
- The final value assigned to a term = 1/level of the word in the synonym tree.
- The terms that follow the term such as “not”, “didn’t”, “doesn’t” etc. are negated.

The final aggregate of all the values of the words is the value assigned to the sentence with respect to the personality parameter.

In order to assign a numerical value to each extracted sentence. The procedure is as follows:
Consider a sentence say,
“He is a great personality as he has helped so many people.”

For this sentence, the synonyms of terms “great”, “personality”, “ helped”, “many”, “people” are fetched from Wordnet. Further, for the synonyms fetched from Wordnet for each word are again subjected to Wordnet for getting synonyms again as shown in fig 4. The synonyms of “great” fetched from Wordnet are significant, huge, famous, noble, wonderful, absolute and important. For simplicity, only one synonym of “great” is considered and again synonyms of “noble” are fetched from Wordnet.

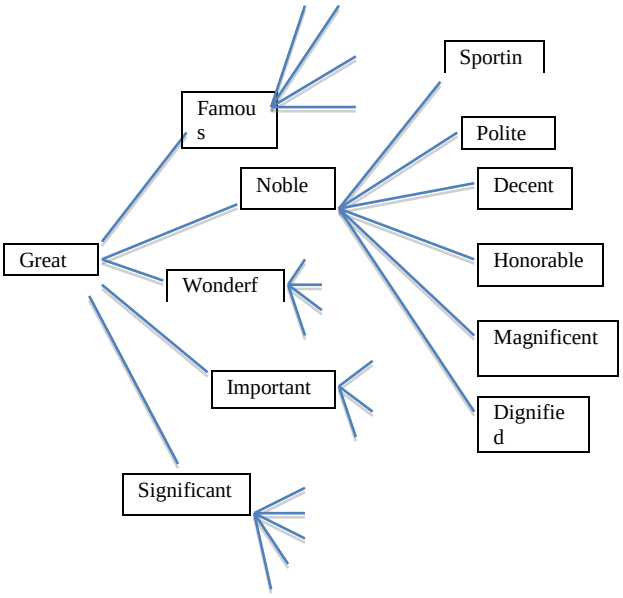

Fig 4. Synonyms of the terms from the Wordnet

The original term “great” is at level 1, “noble” at level 2 and synonyms of “noble” at level 3. Similarly, for the terms significant, huge, famous etc. synonyms are acquired from Wordnet as shown in table III. This process continues to two levels and then all the terms constitute a Termset.

For the example considered, the synonym tree termset would be

T = {great, significant, huge, famous, noble, wonderful, absolute, important, sporting, polite, decent, honorable, magnificent, dignified, valuable,……}

The set T would contain all the synonyms of all the terms in the sentence up to second level as shown in table III. No repetition of the terms is allowed, the set T will have all distinct terms.

TABLE III Terms (words) of the Sentence at Different Levels.

S.no	Terms at level 1	Terms at level 2	Terms at level 3
1	great	significant	
		huge	Vast, gigantic, enormous, massive, giant, titanic
		famous	------
		noble	Sporting, polite, decent, honorable, magnificent, dignified
		wonderful	------
		absolute	-------
		important	Significant, valuable, vital, central, key, essential
2.	Personality	----	----
3.	Helped	-----	-----
4.	Many	-----	----
5.	People	-----	-----

The total number of terms corresponding to each term at level 1 is considered. For example, for “great”, the total number of terms for it is 52 (say 6 synonyms of “ significant”, 6 synonyms of “huge”, 9 synonyms of “famous”, 6 synonyms of “ noble”, 5 for “ wonderful”, 6 for “absolute” and 6 for “ important”). So, total comes to 52 including “great”. In these, 52 terms no term is redundant. Let T be the term set containing all terms from the synonym tree for a given term. Also, let P_i be the term set corresponding to each psychological parameter i, $1 \le i \le 5$.

Eg. P_1= {Honesty, Trustworthy, Fond, True, happy, Soft, Sweet, Warm, Like, Concern, Loving, Supportive, Faith}

P_2= {Love, Dear, Share, Darling, Protect, Good, Provide, Help, Friendly, Thought, Attentive, Pity, Support} and so on till P_5.

Intersection of T & P_i computed i.e. $T \bigcap P_i$.

If $T \bigcap P_i = \varnothing$, then $V_i(P_{i_k}) = 0$, where $V_i(P_{i_k})$ denotes the value associated with i[th] parameter for k[th] term.

If $T \bigcap P_i \neq \varnothing$, then $V_i(P_{i_k})$ = 1/ level at which the term matches.

Therefore, the final value associated with P_i ie $V(P_i) = \sum_{k=1}^{l} V_i(P_{i_k})$, where l is the no. of terms.

So, these final values of $V(P_i)$ for $1 \leq i \leq 5$ are fed to neural networks for classification.

IV. Experiment and Analysis

Tweets of a famous Indian Cricketer (fig 5):
What an experience it was to wave the chequered flag!!! Got to keep it as well!!! :-). Wonderfully organized F1 event by Jaypee. A world class track with excellent facilities for spectators. Truly a memorable day for all of us. One of the happiest moments of my life today. My coach Achrekar sir came to my new house and blessed it with his presence. Happy and safe Diwali to everyone. May God bless our lives with good health and happiness!

Fig. 5. Tweets Extracted from a famous and popular Indian Cricket Player

In the above text, the words with positive intention are memorable, happiest, happiness, etc. these words contribute to the positive effect of the output. Then the ratio of the number of positive attributes and the total number of words are identified. We check for words like not, don't, etc. to find out if there are any words that would negate the meaning of intention of the sentence. Then the same steps are performed for all the negative words. These numbers are calculated for all the personality traits in the system and the resulting values are given as input to the neural network. The neural network classifies it in either positive or negative or neutral.

Fig. 6. Tweets Extracted from one of the authors of the paper

Similarly when tweets of one of the authors (fig 6) of the paper are extracted, he is classified as a neutral person. He in fact is known in his circle as neither a positive nor a negative personality. If the tweets are not sufficient to analyze a person then the message is displayed as " insufficient data for analysis".
Many other known persons are put to analysis using this system. The results are quite encouraging.

V. Conclusion

Text plays a fundamental and the most important role in determining the behavior patterns of any person and the neural network is trained with the training data set to classify the record. As a result, text was extracted from social networking site where most of the common people report their day- to -day activities and they exchange words, here the networking site considered is Twitter. This extracted text is the input to the very important phase of the proposed framework called the Text Preprocessing. The processed text is then converted to numerical values and that is fed to the Neural Networks to obtain the output. Finally we will come to know the behavior of a Human.

The limitation of the application is the usage of Proper and Complete English by the person who will be allied to this application. In this system, we have considered a simple structure of the Neural Network to train the data records. Since accuracy depends on the structure of the network, in future recurrent neural networks will be used. Multi-class labels can be considered for analyzing the behavior of a person. Text can be extracted from other popular networking sites such as Facebook, Google+ and other. Also, currently comparative study of the system implementing clustering techniques for analyzing human behavior with the system based on neural networks is in progress.

Acknowledgement

We would like to thank Dr. J Surya Prasad, Director PESIT, Bangalore South Campus for his constant support and suggestions for carrying out this project.

References

[1] L. Fauset , "Fundamentals of neural networks" Prentice Hall, ed.1, 1993
[2] T.W. Miller "Data & Text Mining A Business Applications Approach" Pearson Prentice Hall, 2008, pp103-124.
[3] S. Bird, E. Klien, E. Loper " Natural Language Processing with Python" Shroff Publishers, ed. 2010.
[4] R. Mayer, "Determining Text Mining with Adaptive Neural Networks" Technical Report for Master's Thesis, February 2004
[5] N.K Nagwani, S. Verma, "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm" in Intl. J. of Computer Applications, Vol 17, no. 2, March 2011.
[6] S. Hariharan, K.R Aashika Rani, "Gender Prediction in Chat based Medium's Using Text Mining ", In Intl. J. of Research and Reviews in Information Sciences, Vol 1, No.1, March 2011.
[7] M. Hearst, "Untangling of Text Data Mining," in the Proc. of the 37th Annual Meeting of the Association for Computational Linguistics,1999(Invited Paper).
[8] M. J.A Berry, G.S. Linoff, " Data Mining Techniques", Wiley Publishing Inc. Second ed. ,pp 240-247.
[9] O. Ozyurt, C. Kose, "Chat Mining: Automatically determination of chat conversations topic in Turkish text based chat based mediums," In J. of Expert Systems with Applications, vol 37(2010), pp 8705-8710.
[10] M. Govindrajan, R.M. Chandrasekaran, "Classifier based Text mining for Neural Networks" In World Academy of Science, Engineering and Technology 27, 2007, pp. 200-207.
[11] B. Yegnanarayana, " Artificial Neural Networks', PHI Learning Pvt Ltd, 2009
[12] www.learnartificialneuralnetworks.com
[13] R. Sharada, A. Arya, S. Ragini, H. Kumar, G. Abinaya, " A text analysis based seamless framework for predicting human personality traits from social networking sites" In Intl. J. of Information Technology and Computer Science (MECS) 2012.

Harish Kumar V is pursuing his post graduate program (MCA) in PES Institute Of Technology Bangalore South Campus, Bangalore, India. Currently doing his final semester internship with Askabt.com. His research areas include text data mining and natural language processing.

Arti Arya has completed BSc(Mathematics Hons) in 1994 and MSc(Mathematics) in 1996 from Delhi University. She has completed her Doctorate of Philosophy in Computer Science Engineering from Faculty of Technology and Engineering from Maharishi Dayanand University, Rohtak, Haryana in 2008. She is working as Professor and Head of MCA dept in PESIT, Bangalore South Campus. She has 15 yrs of experience in academics, of which 7 yrs is of research. Her areas of interest include spatial data mining, knowledge based systems, text mining, unstructured data management, knowledge based systems, machine learning, artificial intelligence, applied numerical methods and biostatistics. She is a life member of CSI and member IEEE. She is on the reviewer board of many reputed International Journals.

Divyalakshmi V has completed MCA from PES Institute Of Technology Bangalore South Campus, Bangalore, India in 2012. Currently she is working with WIPRO. Her major areas of interest are databases, big data and data mining, applications of artificial neural networks.

Nishanth H S has completed MCA from PES Institute Of Technology Bangalore South Campus, Bangalore, India in 2012. Currently he is also working with WIPRO. His interest areas are distributed databases, text mining, web mining.

