A Proposed System for Sentiments Prediction through Social Media Data Analytics

Prof. Swati Chandurkar[1] Himani Asrani[2] Tejal Daga[3] Aanchal S[4] Neha Awate[5]
[1,2,3,4,5]Pimpri Chinchwad College of Engineering, Pune, India

Abstract— Nowadays people are very active and open about expressing themselves on social media due to the liberty provided there. People express all kinds of emotions honestly. Sentiment Analysis is the process of identifying and categorizing opinions and views expressed by humans in a piece of data. This is useful to gain an overview of people and their thought process. This analysis can be put to use in many areas like entertainment, sports, health care, etc. In our implemented system we extract tweets using Twitter API and python functions. We create a dataset from these tweets and give them to Knime tool as training and test dataset. Using different Sentiment Analysis algorithms like Naive Bayes, Decision Tree, etc. the Knime tool gives us output as classified data. The data is classified into 2 polarities: Positive and Negative. The tweets containing words such as happy, excited, great, etc. are assigned with a positive polarity while those having words like sad, sick, worried, etc. are assigned with a negative polarity. Using these polarities, we can predict the moods of the sports players and help improve them so that it doesn't affect the performance. We can use it for movie reviews and also for feedbacks for services of hospitals.

Key words: Twitter, Sentiment Analysis, Polarity, Knime

I. INTRODUCTION

Sentiment Analysis is the process of identifying and categorizing opinions or views expressed by humans in a piece of data.

It is a way to evaluate written or spoken language to determine if the expression is favorable, unfavorable, or neutral, and to what degree. Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation affective state. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.

Data Analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics.

The system aims show how sentimental analysis can help improve the user experience. Using such an analysed data we can also suggest means through which the social media posts and news can be used to improve the mood of the public. This data can also act as a feedback various hospitals, pharmaceutical industries etc. so that they can improvise and meet user's expectations.

We have divided the system in 4 modules: Healthcare, Sports, Entertainment, Emoticons. Different Datasets according to the modules will be given to the system and the Analysis will be done respectively.

II. LITERATURE SURVEY

GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion published by Jonathan Gratch, Gale Lucas and Nikolaos Malandrakis, stated that, Sporting events evoke strong emotions amongst fans and thus act as natural laboratories to explore emotions and how they unfold in the wild. Computational tools, such as sentiment analysis, provide new ways to examine such dynamic emotional processes. In this article we use sentiment analysis to examine tweets posted during 2014 World Cup. Such analysis gives insight into how people respond to highly emotional events, and how these emotions are shaped by contextual factors, such as prior expectations, and how these emotions change as events unfold over time. Here we report on some preliminary analysis of a World Cup twitter corpus using sentiment analysis techniques. We show these tools can give new insights into existing theories of what makes a sporting match exciting. This analysis seems to suggest that, contrary to assumptions in sports economics, excitement relates to expressions of negative emotion. We also discuss some challenges that such data present for existing sentiment analysis techniques and discuss future analysis. They considered the relationship between the proportion of positive, negative or neutral tweets and our other variables of interest: volume (tweets per minute) and unexpected certainty of outcome (the extent to which games are predicted to be "close" (based on betting odds), but ended up with a bigger difference between the teams' scores than expected.

Measuring NBA Players' Mood by Mining Athlete-Generated Content published by Chenyan Xu and Yang Yu stated that online athlete-generated content in social media has high potential to become the information source for both team managers and coaches to discern players' mood status and shaky performance before games. In the existing literature, either in psychology or sport analytics, there is a stream of research that investigated the relationship between athletes' mood and the individual sport performance; however, few of them discussed the causality from the social media perspective. In this study, we look deep into the Athlete-generated content (AGC) and aim to provide a more comprehensive framework to sport operators that incorporates players' social media content into their administrative decision-making process. We obtained a unique and extensive dataset of AGC for active NBA players (in the 2012-13 season) from Twitter and apply sentiment analysis technique to measure the general mood polarity of a player. The general mood was then incorporated into econometrics models to examine its effect on players' individual game performance. The results suggest that the mood of NBA player has significant effect on driving sport performance.

III. ALGORITHMIC SURVEY

Different Sentiment Analysis Algorithms can be used in this System to analyze the data:

A. *Naive Bayes*

Naive Bayes classifier algorithm assumes that the presence of a particular feature in an area is unrelated to the presence of any other feature. It also helps predict the probability of occurrence of a particular string in a given region as compared to the probability of occurrence of another string. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter.

Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

B. *Support Vector Machine (SVM):*

SVM is a supervised machine learning algorithm which is mostly used for data classification problem.

In this, each data item is plotted as a point in n-dimensional space where value f each feature is the value of the coordinate.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

C. *Decision Tree*

Decision tree builds classification or regression models in the form of a tree structure.

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.The topmost node in the tree is the root node.

A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arc leads to a subordinate decision node on a different input feature.
Each leaf of the tree is labeled with a class or a probability distribution over the classes.

D. *Clustering*

It is the task of grouping a set of objects in such a way that objects in the same group are similar to each other.
The objects in a group are dissimilar to objects in another group.

These groups are called Clusters.

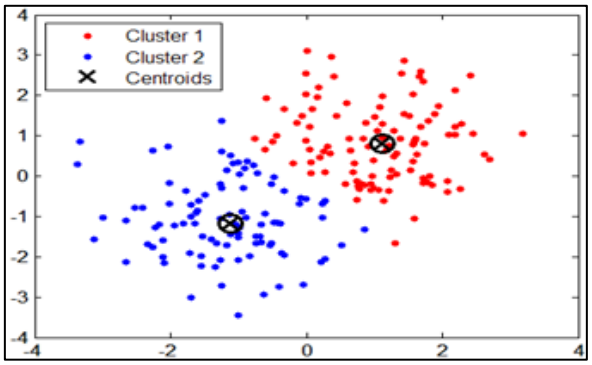

Fig. 1:

The k-means clustering algorithm is known to be efficient in clustering large data sets and is one of the simplest machine learning algorithms.

It aims to partition n objects into k clusters in which each object belongs to a cluster with the nearest mean.

IV. EXISTING METHODS

A. *SentiWordNet*

It is a tool that is widely used in opinion mining, and is based on an English lexical dictionary called WordNet. This lexical dictionary groups adjectives, nouns, verbs and other grammatical classes into synonym sets called synsets. SentiWordNet associates three scores with synset to indicate the sentiment of the text: positive, negative, and objective. The scores, in the values of [0, 1] and add up to 1, are obtained using a semi-supervised machine learning method. For example, suppose that a given synset s = [bad, wicked, terrible] has been extracted from a tweet. SentiWordNet then will give scores of 0.0 for positive, 0.850 for negative, and 0.150 for objective sentiments, respectively. SentiWordNet was evaluated with a labeled lexicon dictionary. To assign polarity based on this method, we considered the average scores of all associated synsets of a given text and consider it positive,if avg score of positive affect is greater than negative affect.

B. *LIWC*

LIWC (Linguistic Inquiry and Word Count) is a text analysis tool that evaluates emotional, cognitive, and structural components of a given text based on the use of a dictionary containing words and their classified categories. In addition to detecting positive and negative affects in a given text, LIWC provides other sets of sentiment categories. For example, the word "agree" belongs to the following word categories: assent, affective, positive emotion, positive feeling, and cognitive process. The LIWC software is commercial and provides optimization options such as allowing users to include customized dictionaries instead of the standard ones.

C. *SentiStrength*

Machine-learning-based methods are suitable for applications that need content-driven or adaptive polarity identification models. Several key classifiers for identifying polarity in OSN data have been proposed in the literature. The most comprehensive work compared a wide range of supervised and unsupervised classification methods,

including simple logistic regression, SVM, J48 classification tree, JRip rule-based classifier, SVM regression, AdaBoost, Decision Table, Multilayer Perception, and Na¨ive Bayes. The core classification of this work relies on the set of words in the LIWC dictionary, and the authors expanded this baseline by adding new features for the OSN context. The features added include a list of negative and positive words, a list of booster words to strengthen (e.g., "very") or weaken (e.g., "somewhat") sentiments, a list of emoticons, and the use of repeated punctuation (e.g., "Cool!!!!") to strengthen sentiments. For evaluation, the authors used labeled text messages from six different Web 2.0 sources, including MySpace, Twitter and YouTube Comments.

D. SASA

We employ one more machine learning-based tool called the SailAil Sentiment Analyzer (SASA) . SASA is a method based on machine learning techniques such as SentiStrengh and was evaluated with 17,000 labeled tweets on the 2012 U.S. Elections. The open source tool was evaluated by the Amazon Mechanical Turk (AMT), where "turkers" were invited to label tweets as positive, negative, neutral, or undefined. We include SASA in particular because it is an open source tool and further because there had been no apple-to-apple comparison of this tool against other methods in the sentiment analysis literature.

V. PROPOSED SYSTEM

We have proposed a system that predicts polarity of a sentence i.e positive or negative. We have divided it in 4 modules, each containing a different extracted dataset. We have extracted the data using Twitter API and Python code specifying the keywords according to the 4 modules. The 4 modules are: 1) Sports 2) Healthcare 3) Entertainment 4) Emoticons. For analysing these datasets we used Knime Tool which is an Open Source Tool. We provide the data in the form of training set and test set to the Knime tool. Different Sentiment Analysis algorithms can be used to predict the polarity of a sentence.

We also plan to analyse the human expressions through Face Recognition System. Sentiments can also be analysed through the voice of a human using Voice Recognition System. For the future, to truly understand and capture the broad range of emotions that humans express as written word, we need a more sophisticated multidimensional scale. Until we are able to measure skepticism, hope, anxiety, or excitement sentiment analysis is (literally) one-dimensional! Sentiments can also be predicted through internal as well as external body movements like heartbeats, dilation of pupils, etc.

Using Sentiment Analysis we can predict the levels of depression that a human faces and thus suggest measures to overcome it.

VI. CONCLUSION

Thus we have studied the concept of sentiment analysis and implemented a system for prediction of sentiments using algorithms like SVM and Naive Bayes.For this we have extracted data from Twitter and used the knime tool.

VII. FUTURE SCOPE

For the future, to truly understand and capture the broad range of emotions that humans express as written word, we need a more sophisticated multidimensional scale. Until we are able to measure skepticism, hope, anxiety, or excitement sentiment analysis is (literally) one-dimensional! Sentiments can also be predicted through internal as well as external body movements like heartbeats, dilation of pupils, etc.

REFERENCES

[1] Marina Boia, Boi Faltings, Claudiu-Cristian Musat, Pearl Pu "A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets" in 2013 International Conference on Social Computing.
[2] Vijay Shankar Gupta & Shruti Kohli "Twitter sentiment analysis in healthcare using hadoop and R." at 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)
[3] Chenyan Xu, Yang Yu "Measuring NBA Players' Mood by Mining Athlete-Generated Content" in 2015 48th Hawaii International Conference on System Sciences.
[4] Georgios Solakidis, Konstantinos "A Multilingual Sentiment Analysis Using Emoticons and Keywords."
[5] Adela Ljajić , Branko Arsic"Sentiment analysis of textual comments in field of sport "
[6] V.K. Singh, R. Piryani, A. Uddin, P. Waila "Sentiment Analysis of Movie Reviews and Blog Posts" in 2013 3rd IEEE International Advance Computing Conference (IACC)
[7] An Introduction to Sentiment Analysis - Ashish Katrekar (AVP, Big Data Analytics)
[8] Opinion mining and sentiment analysis - Bo Pang and Lillian Lee
[9] http://www.stefanoscerra.it/movie-reviews-classification/\
[10] http://adilmoujahid.com/posts/2014/07/twitter-analytics/

