**IN depth Opinion mining and sentiment analysis of twitter data**

**Eashan Sharma eashansharma31*@*gmail.com Aryan gaur aaryangaur06*@*gmail.com**

**Manav Rachna International Institute of Research and studies**

***ABSTRACT*:-**In today’s high-tech world opinion mining and sentiment analysis has become a major topic which is growing day by day and is utilized in many real world applications. Nowadays instead of gathering feedback from friends and family for purchasing a specific item, we determine the opinions of various Individuals across the world by microblogging data. A very popular social media platform where we can express our opinions and interact with other people is twitter. The tweets posted on twitter based on their emotional content can be categorized into positive, negative or irrelevant with the help of various classifying algorithms. In this paper we train a sentiment classifier which can help us in accomplishing various tasks.

**I: Introduction** “Sentiment analysis additionally called opinion mining or feeling AI refers to the employment of linguistic communication process, text analysis, linguistics, and bioscience to consistently determine, extract, quantify, and study emotive states and subjective info. Sentiment analysis is wide applied to voice of the client materials like reviews and survey responses, on-line and social media, and tending materials for applications that vary from selling to client service to clinical medicine”**[14].**Twitter is among most widely used social media platform to express their thoughts and emotions whilst interacting with other people online. **[1**] Many users use twitter to convey their opinions and thoughts on different topics, events, different reviews on products, services, and other Twitter users they are interested in. Sentiment analysis on such tweets can be performed by different firms to get an insight on peoples’ thought process on different events to predict people’ response towards their products. Thus, it is evident that interpreting user perception from Twitter data is very useful for many applications. “The main objective of this research, using a case study of FIFA World Cup 2014, is to examine the adequacy of different famous classifiers and recognize the more reasonable classifier(s) for Twitter that could facilitate the way toward arranging estimations in tweets" [15]. "In this examination we will utilize the previously mentioned contextual analysis to see and find out about a noticeably utilized book arrangement strategy called Bayesian Logistic Regression (BLR) for notion order i.e, giving positive or negative supposition on tweets and [16] distinguishing classifiers with a satisfactory exhibition that could be utilized to group tweets dependent on the communicated feeling as unbiased, polar and superfluous; and afterward polar into either good or negative".

On Twitter it is not an exaggeration that people tweet about anything and everything thus, the most basic method to mine twitter sentimental is reading data and analyzing and hashtags (#) as discussed in section 3. section 4 discusses the approaches used for classification, section 5 discusses salient and various features of classifiers, sections 6 deals with various ML classifying algorithms, section 7 is a case study of 2014 FIFA cup and section 8 and 9 show conclusion and references respectively.

**II: LITERATURE REVIEW** We have extracted, analyzed and combined data from some of the most prominent research papers like from “Twitter sentiment Analysis” by Theresa Wilson and Johanna Moore we have extracted the data used for the training the classifier. From “Opinion mining and sentiment Analysis on twitter” by balkrishnan and gokulkrishnan ,we have made use of the various preprocessing techniques and discussed various classifying algorithms and their performance. From “Opinion mining and sentiment polarity of twitter” by Peiman barnagi and John breslin we have made use of a case study of 2014 FIFA world cup to find relation between events and twitter sentiment by using Bayesian logistic regression**.**

**III: TWEET SENTIMENT ANALYSIS METHOD**

Given below is an opinion mining model on twitter data which is shown as a block diagram. In this various blocks represents different steps of preprocessing like feature extraction, filtering and so on. There is so much useless information in tweets for which a workflow is designed. In this comes cleaning using tokenization, filtering, decrypting the content of messages such as username, URL’s to general tags, stemming and uppercase conversion and hashtag direction to mark the topics and keywords. Then finally we use a trained model to find relations between tweets and events that occur.

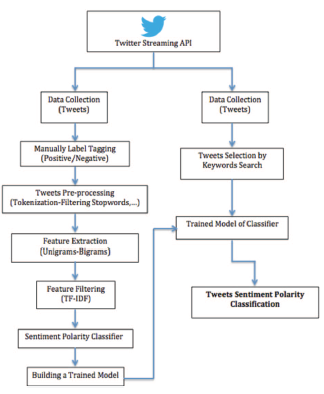
****

Fig. 1. Trained model to check polarity by applying it on some tweets

**IV: Case Study of 2014 FIFA**

“In 2014 world cup 672 million tweets were posted after 64 football matches” [1]. To find

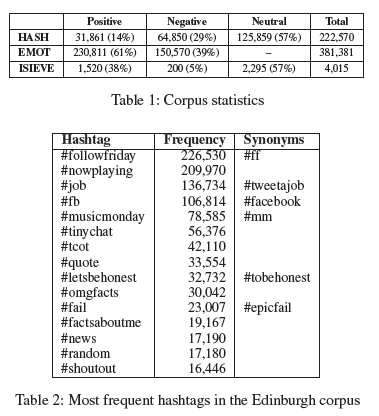
Relations between the sentiment and major events of 2014 world cup, we have used extracted twitter sentiment in this paper. We use of twitter’s streaming API for mining tweets and filtering them using some of the official world cup hashtags. In this paper we look at major talking events of the tournament. There were two major incidents, firstly a Uruguayan player accused of biting an Italian player and secondly the elimination of the host country of world cup brazil. After this a flood of positive and negative tweets were received which showed changing sentiments based on events and timestamps.

**V: Tweet Collection**

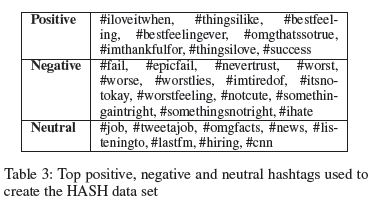
To build a model data was gathered in two ways the first was used to create a training set by using data from twitter’s streaming API. The tweets were labeled as “negative” or “positive” manually .Then secondly we mine tweets relate to the world cup by filtering the tweets using the official hastags that were used and also the hashtags of the teams as well (e.g.“#ARG” and “#GER”). Players usernames also helped in extracting tweets (e.g. “#suarez” and “#brazil”).The data was created in JSON format as aset of document. [2].

**VI: Training Dataset**

Many different corpora of twitter messages is included in our paper. They are hastagged data set(HASH) used for development and training compiled from Edinburgh data set(EMOT).

****

**1: Hashtagged data set**

A subset of Edinburgh twitter corpus is the hashtagged data set. Over a period of two months 97 million tweets were collected. By filtering duplicate tweets, non-english tweets and tweets without hashtags is how we create hastagged data set. Identification of positive negative or neutral messages is done from the remaining set(about 4 million). For training and development we use these hashtags. Some common Tweets like: #fail, #worst, #iloveitwhen, etc indicate polarity. We have shortlisted the top hashtags that appeared a 1000 times to identify positive, negative and neutral tweets. These hashtags and their polarity are given in Table 3.****

**2: Emoticon data set**

Go bhayani and huang created the emoticon data set for a project at standford by collecting positive ‘:)’ and negative ‘:(’ emoticons. Messages that contained both of them were removed. We have only utilized their training data set and not the dataset for evaluation.

**VII: DATA PREPROCESSING**

We should use preprocessing techniques to assess certain tokens of tweets as it is very likely that tweets may contain spelling or grammatical mistakes, acronyms, colloquialism and slang because twitter has incorporated 140 character limit on tweets. The irrelevant content is filtered out and only the relevant remains due to preprocessing techniques. Sentiment analysis in micro blogging utilizes these techniques for many information retrieval applications. Now a series of preprocessors assist in conversion of the message strings into feature vector by passing the collected data through it. Some of the preprocessing techniques are explained below. The performance of the classifier is dependent on the quality of features/attribute extracted by said data preprocessing techniques.

**1: Tokenization**

First step of preprocessing, splits the string into tokens and construction of bag of words. To form individual words , splitting of text with white spaces is done. Sentiment classifier can be trained using tokens which is a word in a sentence.

**2: Removing Stop-Words**

Articles and prepositions are frequently encountered in texts and are known as stop words. As they have no vital part in final sentiment of the text they can be removed from the bag of words .Dictionaries are used against the text to check every word. Stop words like “but”, “or”, “until”, “also”, “able”, “and”, “as”, are included. Matching elements are also removed.

**3: Twitter Symbols**

Words after the image "@" are utilized for usernames and "#" is utilized to check points or catchphrases in a tweet."The change of usernames and URLs to nonexclusive labels is being done and a few notices are additionally used to improve the presentation of the supposition classifier" [3].

**4: Damming**

To reduce different forms of words like nouns, verbs, adjectives etc affixes are removed from the words and are replaced with the root to get a common base form. This technique is called damming or stemming. (e.g. the words “synthesis”, “synthesized” and all other types of this word are converted to “synthesi” after damming).“Different Weka packages are used for this operation. The output of sentiment classifier is improved and dimensionality of bag of words is reduced” [4].

**5: Replacing Emoticons**

To express emotions in a concise manner emoticons are used by users in various posts. It is an easy way to determine the polarity of a message. SMILE or FROWN keyword can be replaced with a bunch of emoticons, like :) :( :D =] :] =)= [ =( .

**6. Identification**

To communicate amazing feelings, capital letters are utilized (e. g. WOW!).It is a decent marker of extremity of message and is called as e-yelling. This preprocessing step extricates this element by recognizing the uppercase words before eliminating packaging.

**7: Lower Casing**

To ensure correct mapping between tokens and features the entire word should be in a consistent case while classifying text. It is not uncommon to find irregular casing in research work such as"AIYsiSOfTwITteRseNtlMeNtan").

**8: URL Extraction**

To give additional content in the limited character post some users post URLs in their tweets. The user might provide supplement information regarding his emotion in a URL posted in a tweet, but it is rather difficult and expensive to crawl URLs for their content. All URLs in the preparation tweets have been supplanted with a proportionality class <URL> in order to tend to down the element size during preparing, The lessen in highlight size is noteworthy.

**VIII: Features of a Trained Classifier**

Various features like unigrams and bigrams are used for classification. Features from sentiment lexicon and POS features used in sentiment analysis are also included. To determine sentiment of the tweet by extracting useful words various methods are used. This section is about feature extraction, selection and external lexicons which compare pre-defined features with the extracted ones.

**A: Feature Extraction**

Feature extraction can be defined as selecting important words as features of a text and removing the remaining words that don’t contribute towards the sentiment of text. Thus obtaining more accurate sentiment is obtained for the tweet and the noise is filtered

**1: Unigram features:**

This is most usually utilized technique for include extraction and is utilized to misuse the requesting of words by looking at each word autonomously, at once in a book, which can be stretched out to N-gram highlight. It very well may be executed in characters, words or sentences.

**2: N-gram features**

When we take a set of sequential words in a set then it is called an N-gram feature. If we take N=2, then we look at two sequential words at a time which is known as a bigram.

“Some works on unigram shows that classification performance is dependent on the kind of dataset used. Sentiment polarity classification on movie reviews yield better performance when unigrams are used( Pang et al)”**[7]**.As the max limit is 140 characters most of the tweets are about 30 characters long , so mainly unigram and bigram features are used for sentiment classification.

**3: Lexicon features**

Words included in the lexicon are tagged with their predefined polarity and are marked positive, negative, or neutral. It is further divides based on the presence of words from the lexicon.

**4: Micro-blogging features**

To capture the presence of negative, neutral and positive emoticons along with abbreviations and intensifiers (e.g., all-caps and character repetitions) binary features are used. Various internet lingo and slang dictionaries are used for this process.

**B: Feature Filtering**

At the point when size of corpora is huge, a noteworthy number of highlights are kept, so we utilize various strategies to choose the top highlights to prepare the classifier."To channel the highlights utilizing the quantity of words in the content by weighting and scoring each of the unigrams and N-grams a mathematical factual strategy Known as Term Frequency-Inverse Document Frequency (TF-IDF) is utilized" **[7].**

**IX: Sentiment Classifier**

We use several different popular classifiers to perform opinion mining of the preprocesses tweet data, so that we may identify the classifiers that are performing better and gauge their performance. Weka, an extremely viable information mining and AI device, created in New Zealand, has been utilized for this reason as it has a great deal of ordering calculations that are habitually utilized in different investigations. Some Of them are clarified underneath:

**1: Bayesian Logistic Regression**

"To perform text arrangement by at the same time choosing the highlights and contracting them BLR model is utilized. It produces exiguous prescient models for text information and furthermore utilizes a Laplace before maintain a strategic distance from over fitting" [5].

**2: Naive Bayes**

“Guileless Bayes classifier is a probabilistic classifier dependent on Bayes hypothesis which computes the likelihood of extremity of tweets, with a suspicion that all highlights are free" [9]."Naive bayes didn't yield more noteworthy outcomes contrasted with different classifiers delineated in this area, despite the fact that it improved outcomes in" [10].

**3: Random Forest**

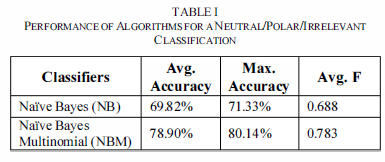
“Random Forest classifier works on method of sub divisions, that is, tree structure” **[11].** The class with highest turns will be chosen which is given by each tree to the input vector.“ The blunders that happen in this model rely upon the quality of each tree in the timberland and furthermore the connection between any two trees in the backwoods. The trees ought to be solid and free of one another to limit blunder rate" [12].

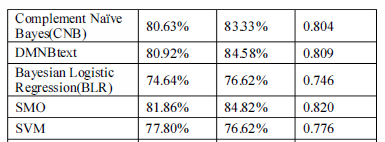
**4: Support Vector Machines (SVMs)**

"Part replacement technique is the base of SVMs class of calculations" [13].Hypothesis space of straight capacities in a high dimensional element space is utilized by the framework. It executes a taking in inclination got from factual learning hypothesis after it is prepared by a learning algorithm."To make profoundly non direct arrangement without stalling out in the nearby minima SVMs are the most ideal choice" [13].

**X: RESULTS AND DISCUSSION**

Table 1 shows performance of polar, neutral and irrelevant classification observed by each classifier.

****

****

Some of the widely used algorithms fail to achieve satisfactory performance which is quite surprising. An average accuracy of about 80% was achieved using the naïve bayes classifier. To get accuracy more than 80% SMO

And Random Forest classifiers should be used in those cases. To get the best average accuracy SMO is the classifier that comes on top.

**XI: SENTIMENT ANALYSIS ON SOME MAJOR WORLD CUP 2014EVENTS**

Open conclusion when taken in gathered form during uncommon occasions, for example, the FIFA world cup is given through tweets.“ The trained model in this paper was used to carry out sentiment analysis for the events which occoured during the tournament”**[6].** As more than 51.56% tweets were in English as shown by the language classifier of the world cup analysis of only English tweets were done from the gathered 30 million. To see extremity esteems for various substances and how because of different occasions they change after some time the extremity estimations of these tweets were utilized.

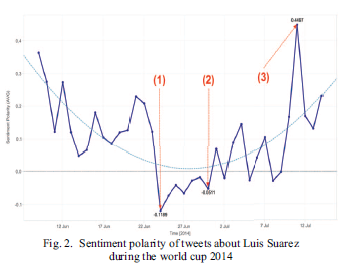
**1: A Major Event during the World Cup**

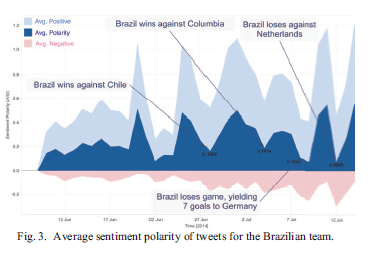
To discover relations between tweets and the events that occurred we extracted all the related tweets of that major event. Luis Suarez, a Uruguayan player was claimed of gnawing an Italian safeguard, Giorgio Chiellini, on June 24th. An unexpected explosion of negative tweets was trailed the occurrence. All the tweets that referenced the player's name, conclusion grouping was performed utilizing the prepared model. (Figure 2) shows the sentiment classification output which is divided into 3 different sections based on the trend of tweet polarity of sentiment for the mentioned player. The polarity values before the biting incident is shown in the first part Player performance and match results is responsible for fluctuation of sentiment polarity rates. The vast majority of the slant is positive with various power, (for example, emphatically certain). The start of a negative pattern after the episode is appeared in the second aspect of the opinion extremity. Generally notion of tweets is negative with various rates. After Suarez gave an open statement of regret on June 30th, it was good for the Twitter people group and results in a positive pattern lastly it arrives at a pinnacle level of positive extremity which is appeared in third aspect of the conclusion extremity when he marked his new agreement with Barcelona FC. Based on the three incidents Fig.2 shows gives graphical representation of same:-

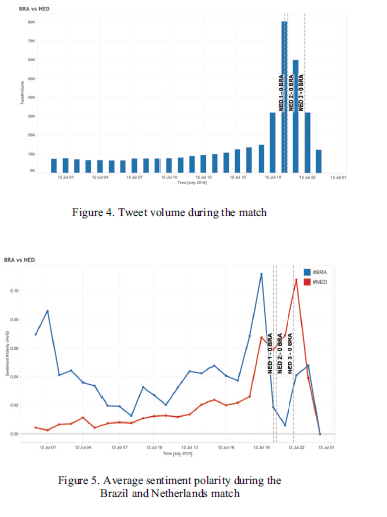
(1)\_ Italy’s defender allegedly bitten by Suarez

(2)\_ Public apology issued by Suarez

(3)\_ Barcelona signs a contract with Suarez

**2: Another Major Event as Elimination of Brazil**

Brazil being eliminated from the tournament was another major event as it is one of the best teams in the world and it also hosted the 2014 FIFA World Cup. Fig3. shows the overall polarity (middle diagram) and the trend of positive or negative tweets which is the base of average sentiment polarity. After each victory or loss in matches polarity of sentiments changed.****

There is a noteworthy difference in extremity from negative to positive subsequent to losing the match against Germany (1-7) and before the third spot end of season game among Netherlands and brazil. Individuals were trusting that the past destruction may draw out the best in the Brazilians last game, however during the third spot play-off the in the wake of yielding three objectives against the Netherlands the bearing changed once more. During the match the expansion in the quantity of tweets posted is appeared in Fig. 4 and during Brazil's last match the extremity changes of slants is appeared in figure 5. ****

**CONCLUSION:-**

We can utilize the sentiment classifier to check the polarity for various entities and events to see if it is perceived as positive or negative in public’s eye. An insight on public’s opinions about a certain event can be known by analyzing the sentiment of tweets. The events can be the FIFA world cup, teams, players and more. Assessments will undoubtedly change over the long haul during a significant occasion or after shameless conduct. The model referenced in this paper was prepared by twitter information by utilizing text highlights to decide the extremity of the supposition for significant occasions that happened during the world cup. The aftereffect of the contextual analysis shows positive and pessimistic response of individuals about the occasions that happened and how it can likewise change after certain episodes during those occasions. This kind of assessment investigation which uses stubborn writings from twitter information encourages us to separate examples which can be utilized to anticipate future occasions.

**REFRENCES**

[1] Barnaghi, P., Ghaffari, P. and Breslin, J., 2016. Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment. *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, [online] Available at: <https://ieeexplore.ieee.org/abstract/document/7474355> [Accessed 4 September 2020].

[2] Terrana, D., Augello, A. and Pilato, G., 2014. Automatic Unsupervised Polarity Detection on a Twitter Data Stream. *2014 IEEE International Conference on Semantic Computing*, [online] Available at: <https://ieeexplore.ieee.org/abstract/document/6882013> [Accessed 4 September 2020].

[3] Zhang, L., 2020. *Sentiment Analysis On Twitter With Stock Price And Significant Keyword Correlation*. [online] Hdl.handle.net. Available at: <http://hdl.handle.net/2152/20057> [Accessed 4 September 2020].

[4] Jivani, A.G., 2011. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), pp.1930-1938.Available at: <https://kenbenoit.net/assets/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf>

[5] Genkin, A., Lewis, D. and Madigan, D., 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, [online] 49(3), pp.291-304. Available at: <https://www.tandfonline.com/doi/abs/10.1198/004017007000000245> [Accessed 4 September 2020].

[6] Barnaghi, P., Ghaffari, P. and Breslin, J.G., 2015, August. Text analysis and sentiment polarity on FIFA world cup 2014 tweets. In *Conference ACM SIGKDD* (Vol. 15, No. 2015, pp. 10-13). Available at:<https://www.researchgate.net/profile/Peiman_Barnaghi2/publication/280715888_Text_Analysis_and_Sentiment_Polarity_on_FIFA_World_Cup_2014_Tweets/links/55c2202d08aeca747d5dc565/Text-Analysis-and-Sentiment-Polarity-on-FIFA-World-Cup-2014-Tweets.pdf>

[7] Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*. Available at: <https://arxiv.org/abs/cs/0205070>

[8] Pandey, V. and Iyer, C., 2009. Sentiment analysis of microblogs. *CS 229: Machine learning final projects*. Available at: <http://cs229.stanford.edu/proj2009/PandeyIyer.pdf>

[9] Murphy, K.P., 2006. Naive bayes classifiers. *University of British Columbia*, *18*, p.60. Available at: <https://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf>

[10] Pak, A. and Paroubek, P., 2010, May. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).Available at: <https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pdf>

[11] Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32. Available at: <https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf>

[12] Breiman, L. and Cutler, A., 2005. Random Forests. Berkeley.. [Apr. 1,2012].

[13] Bennett, K.P. and Campbell, C., 2000. Support vector machines: hype or hallelujah?. *Acm Sigkdd Explorations Newsletter*, *2*(2), pp.1-13. Available at: <https://dl.acm.org/doi/abs/10.1145/380995.380999>

[14] Pang, B. and Lee, L., 2009. Opinion mining and sentiment analysis. *Comput. Linguist*, *35*(2), pp.311-312. Available at: <https://www.aclweb.org/anthology/J09-2007.pdf>

[15] Bifet, A. and Frank, E., 2010, October. Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science* (pp. 1-15). Springer, Berlin, Heidelberg. Available at: <https://www.cs.waikato.ac.nz/ml/publications/2010/Twitter-crc.pdf>

[16] Mishne, G., 2005, August. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access* (Vol. 19, pp. 321-327). Available at: <https://pdfs.semanticscholar.org/0982/8f26fd9bb7ef105538fa51a57456ae38e63e.pdf>