SCIREA Journal of Computer
http://www.scirea.org/journal/Computer
May 8, 2022
Volume 7, Issue 1, February 2022
https://doi.org/10.54647/computer52271

Machine Learning Techniques for Opinion Mining from Social Media

K. Victor Rajan, Freddy Frejus

Department of Computer Engineering, Atlantic International University, Honolulu, Hawaii 96813, USA

*Corresponding author: victor@jts.co.in

Abstract

Expressing emotions through various channels is part of human life. Directly or indirectly, we somehow reflect our opinions through speech, writings, etc., in our daily life. Opinions containing emotional or sentimental words have huge impact in the society. Analyzing the emotions and sentiments of people has its own importance. For example, we can measure the well being of a society, prevent suicides, and measure the degree of satisfaction of their customers by analyzing the comments or the feedback. The world wide web sites like social media, forums, review sites, and blogs generate a large volume of data in the form of opinion, emotion, and sentiment about social events, government policies, political events etc. Increased use of technology has made people proactively express their opinion through social media sites like Twitter, Facebook, and Instagram. Decision makers can make use of social media content to understand how people react to policies, events, and consumer products. But, social media analytics is a complex task due to the challenges in the natural language processing of social media language. These messages do not adhere to grammatical standards. The unstructured data from the social media needs to be cleansed and well-structured for opinion mining. These messages often reflect the opinion, emotion, and sentiment of the public through a mixture of text, emoticons, image, etc. Standard natural language processing tools cannot be used to analyze the emotion or sentiment hidden in the social media content. Social media users use emoticons like smiling face (☺), angry face (☹) etc., to express emotion instead of words. They also express positive (👍) or negative (👎) sentiment using emoticons. These statements are called electronic Word of Mouth (eWOM) and are much popular in business and service industry to enable customers to express their point of view. We propose to use a two-step approach for opinion mining of social media content. Instead of using language parsers for parsing the eWOM, we propose to use machine learning algorithm for opinion mining.

Keywords: Emotion Analysis, Sentiment Analysis, Opinion Mining, Social Media Analytics.

1 INTRODUCTION

People convey their opinion in social media sites using short messages. These messages often reflect their emotions on public policies, politics, etc. Analyzing the public opinion from the textual data over the internet will add commercial value to business and government organizations. For example, we can measure the degree of satisfaction of customers of a product. This can quickly alert product owners when customer preferences and desires change. Opinion mining systems use natural language processing and text analysis to determine the impactful opinions hidden in a particular text. Opinions containing neutral judgments express facts and truths. However, the emotion and sentiment analysis is the basic idea behind opinion mining for business organizations. Text based emotion analysis introduces some challenges in our work in the sense that social media users do not follow good communication style. The messages are not written as per grammatical standards. They contain the feelings of individuals through mixture of symbols, text, emoticons etc. The unique characteristics of eWOMs are; short and noisy content, large data volume, diverse and fast changing topics, etc. The research on social media analytics is still evolving. It may not be possible to write natural language parsers to understand the hidden context of these messages. Impactful opinions are mainly expressed using emoticons and symbols. The challenges involved in opinion meaning of eWOMs are:

- Identifying the type of emotion expressed.

- Identifying the sentiment polarity.
- Predicting the future growth of a topic.

The first two are classification problems. The third is a prediction problem. Unless we solve the above, it is difficult to extract the opinion successfully. The computational approaches include all the techniques that are needed to design and implement an opinion mining system. Computational approaches involve collecting a data set and applying systematic algorithms to detect the hidden opinion. Emotion and sentiment analysis algorithms can be classified into two major types as follows:

i) **Lexicon based**: An emotion or sentiment lexicon is a knowledge repository containing textual units annotated with labels. They rely on lexical resources like lexicons, bags of words or ontology.

ii) **Machine Learning based**: The machine learning approach uses artificial intelligence and machine learning algorithms that can learn from data by making use of document similarity among text messages.

Lexicon based algorithms are not suitable for parsing the eWOMs since these messages are unstructured. We propose an approach that focuses on identifying the type of emotion or sentiment using machine learning algorithm. Machine learning classification models are used to assign data to a discrete group based on a specific set of features. Emotions and sentiment are classified using trained ML models.

Prediction of growth or decline of an event using historical data can be done using the following approaches:

i) **Statistics based**: Statistical analysis is the process of understanding current trend in order to predict future values using time series, moving averages, etc.

ii) **Machine Learning based**: The machine learning approach relies on artificial intelligence and machine learning algorithms to predict the future values by making use of linear or stochastic models.

The recent research in machine learning shows that ML algorithms are broadly divided into supervised and unsupervised learning methods. Mixture of text, emoticons, and hash tags is a new style of writing and conventional approaches fail to recognize eWOMs. We propose an approach that performs opinion mining of eWOMs as follows:

1. Focus on identifying six major discrete classes (Happy, Anger, Sad, Fear, Disgust, and Surprise) of emotion.

2. Identify sentiment polarity (Positive or Negative) hidden in the message.

3. Measure the momentum of topic using growth of hash tags.

Now, opinion mining becomes classification and prediction problem. Efficient classification and prediction models are available in artificial intelligence and we can use machine learning to solve our problem efficiently.

The remainder of this paper is organized as follows. In section 2, we present our research methodology and the architecture of opinion mining system. In section 3, section 4, and section 5, we discuss the machine algorithms suitable for emotion analysis, sentiment analysis, and prediction of growth respectively. Finally, the paper is concluded in Section 6.

2 RESEARCH METHODOLOGY

Understanding the language of eWOM by computer systems is a complex task due to the non-standard structure of messages. We cannot use standard natural language processing tools to analyze the emotion or sentiment**.** Our model for analyzing emotion is tailored to handle the style and specifics of this informal writing culture. The motivation behind our research is to improve social connectivity and emotional expressiveness of real-time messaging. In order to identify emotion in eWOM, our model processes symbolic cues, such as emojis, transforms them to words, and then employs machine learning techniques to classify the emotion category. For example, the sentence 'Enjoying my lazy Sunday ☺ ' represents a happy message that does not contain the word 'Happy'. Because, the word 'enjoy' and the emoji ☺ express happiness. Opinion mining will not be accurate unless we translate these messages into plain text without losing the context and emotion attached. The real context of the message will be lost unless we take the emoticons into consideration. So, we propose a two-step model to process the eWOMs. Our first step is to develop a feature extractor that will produce a plain sentence from eWOM. The Feature Extractor (FE) works as follows.

- Pick up only English messages. Drop other language messages.

- Replace emojis with CLDR meaning if mapping is available (Table 1). Remove otherwise.

- Convert all words are to lowercase.
- Remove stop words and punctuation like periods, commas, brackets, etc.
- Remove all words not purely comprised of alphabetical characters (words containing special characters and numbers).

Following table shows examples of emojis translated using Unicode Common Locale Data Repository (CLDR).

Table 1: CLDR mapping of Emotions.

UNICODE	EMOJI	CLDR MEANING
U+1F600	😀	Grinning face
U+1F642	😊	Smiling face
U+1F609	😉	Winking face
U+1F615	😕	Confused face
U+1F622	😢	Crying face
U+1F602	😂	Tears of joy
U+270C	✌	Victory hand

The following table shows examples of eWOMs translated using our feature extractor.

Table 2: Sentence Formation using Feature Extractor

eWOM	ENGLISH SENTENCE
Salute to our warriors 😂 who taught us to raise voice against evil!	Salute to our warriors with tears of joy who taught us to raise voice against evil
Enjoying my lazy Sunday😊!!	Enjoying my lazy Sunday with smiling face
Samsung Galaxy M12 is out ✌!	Samsung Galaxy M12 is out with victory hand

The translated sentence captures the sentiment without losing the original emotion. Following diagram shows the architecture of our opinion mining system.

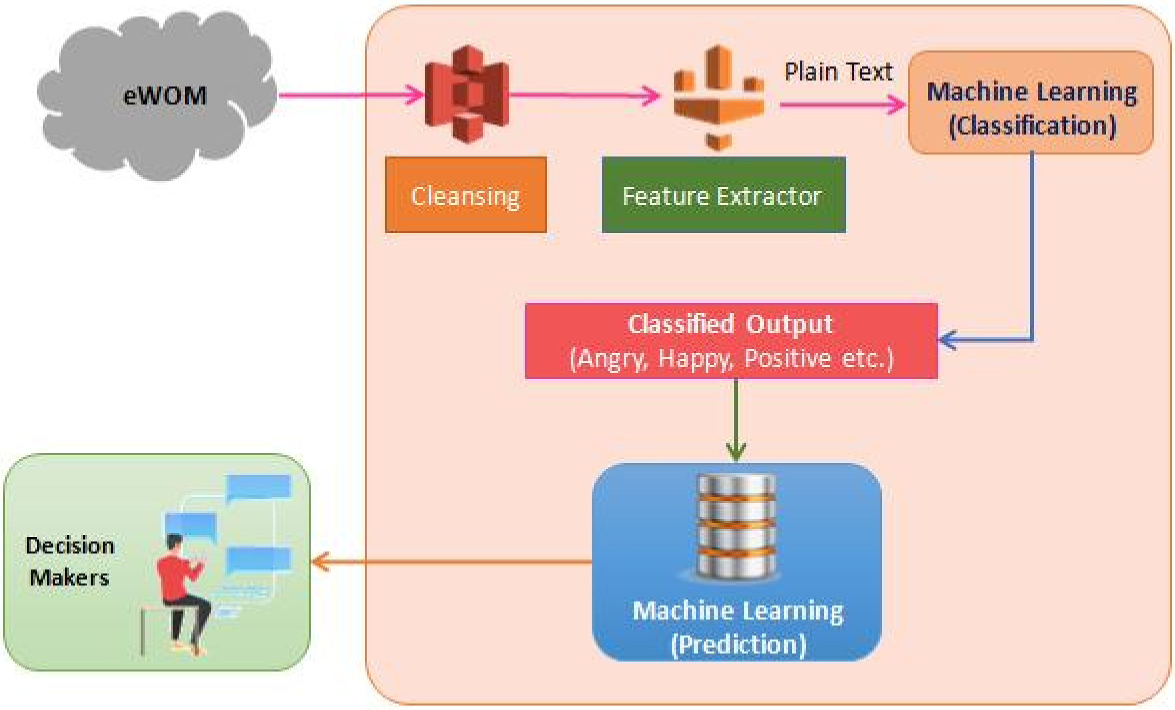

Figure 1: Architecture of Opinion Mining System

3 EMOTION ANALYSIS USING SUPERVISED LEARNING

Feature Extractor converts the eWOM to plain text. Researchers identified two major approaches to analyze the emotion from text messages, namely emotional categories and emotional dimensions. The emotional categories approach divides the emotions into discrete emotion labels, one of its notable works presented by Cecilia et al [1]. Neviarouskaya et al. [2] used a rule-based method for determining Ekman's basic emotions in blog posts. The dimensions approach represents the emotion classes in a 2D or 3D dimensional form. Each emotion occupies a distinct position in space. Bradley [3] proposed a 2D approach that uses two vectors pointing in two directions assuming the presence of an arousal dimension with valence dimension vector determining the direction in which a particular emotion lies. Plutchik [4] proposes a 3D model arranging the emotions into concentric circles with inner being the basic emotion and the outer more complex emotion. The emotion category approach is suitable for our research since our goal is to identify six major types of emotions (Table 3). We use supervised learning ML algorithm for classification. Our model learns to label a given input (sentence) to the corresponding output (emotion category) based on the test samples used for training. Pairs of sentence vectors and labels (e.g., Happy, Anger) are fed into the machine learning algorithm to generate a model. The well-trained model generates predicted label (Happy, Anger, etc.) for a given input.

3.1 EMOTION CLASSIFICATION USING K-NEAREST NEIGHBOUR (KNN) ALGORITHM

KNN is a popular and widely used machine learning algorithm for text classification. This algorithm classifies input into one of the predefined categories of a group that was created by supervised learning. Our training data set consists of six categories as mentioned in Table 3. KNN algorithm classifies a document by looking at the training documents that are most similar to it. The algorithm assumes that it is possible to map the documents as points in the Euclidean space. Euclidean distance is the linear distance between two points in Euclidean space. The distance between two points in the space with coordinates p=(x, y) and q=(a, b) is calculated using the formula

$$d(p,q) = d(q,p) = \sqrt{(x-a)^2 + (y-b)^2}$$

Hence, identifying a suitable method to calculate the Euclidean distance between two text messages is crucial and important for the success of our algorithm. We use Term Frequency-Inverse Document Frequency (TF-IDF) method to find the similarity between two documents. TF-IDF method determines the relative frequency of words in a specific document through an inverse proportion of the word over the entire document corpus. TF-IDF is the product of the TF and IDF scores of the word. In TF-IDF, similar text must result in closer vector.

TF = number of times the word appears in the doc/total number of words in the document.

$$f_{ij} = \textit{frequency of term } i \textit{ in document } j$$

IDF = ln (total number of docs/number of docs the word appears in)

$$idf_i = log_2\left(\frac{N}{df_i}\right)$$

Hence,

$$\text{TF-IDF} = tf_{ij}idf_i = tf_{ij} \times log_2\left(\frac{N}{df_i}\right)$$

Higher the TF-IDF score, the rarer the term is and vice-versa. Smaller Euclidean distance between the documents indicates their higher similarity. Distance 0 means that the documents are completely equal. KNN classification algorithm combined with TF-IDF for distance calculation during our experiments yielded prediction accuracy of more than 80%.

3.2 EXPERIMENTS AND RESULTS

We used Twitter data related to product reviews and social events for our experiments. In this section, we describe the data sets used, the evaluation tasks, and the experimental results of our approach.

3.2.1 Training Data Set

We created a data set from twitter social media for supervised learning. Our experts collected around 13,000 tweets, cleansed, and classified them manually into six categories. In English, more than one word may convey the same emotion. Words synonymous with emotions like happy, anger, etc., were used by experts as shown in the below table during our classification.

Table 3: Dictionary for Emotion Lookup

No	Emotion Category	Synonyms in Dictionary	Count
1	Happy	Joy, Smile, Laugh, Enjoy, Cheer, Glad	3524
2	Anger	Ballistic, Furious, Outraged, Infuriate	2543
3	Fear	Terror, Horror, Unease, Worry, Anxiety	1987
4	Sad	Sorry, Regret, Depress, Dejected	2405
5	Disgust	Dislike, Revolt, Objection, Hatred, Repel	1104
6	Surprise	Astonish, Amaze, Marvel, Wonder, Shock	1346
		Total	**12909**

Though, impactful emotions are few (six in our experiment), not all people use the same word to express a particular emotion. Our approach of using synonyms increases the breadth of coverage and hence improves the efficiency of our ML algorithm. The hand crafting also helps us to be highly subjective and capture context sensitive information. For example, a message containing ‘Don’t worry’ will be classified as ‘Happy’ and not ‘Fear’.

3.2.2 Empirical Results

Twitter API provides options to search for messages using filters like place, language, etc. Our data for evaluation was a large set of tweets from Sep. 2021 to Nov. 2021 in the English language.

Table 4: Twitter Data Set for Experiment

DESCRIPTION	SEP 2021	OCT 2021	NOV 2021
Total number of tweets	5,234,667	5,201,745	5,134,443
Tweets with emotion	1,465,706 (28%)	1,352,454 (26%)	1,386,299 (27%)
Tweets with emotion and hashtag	628,160 (12%)	572,191 (11%)	616,133 (12%)

From the above table, we observe that enough number of tweets with emotion are available for analysis. Our experiments were conducted with values for k = 3, 5, 7, 9, 11 and k = 9 produced high accuracy. The below figure shows the knn-classification done during our experiment.

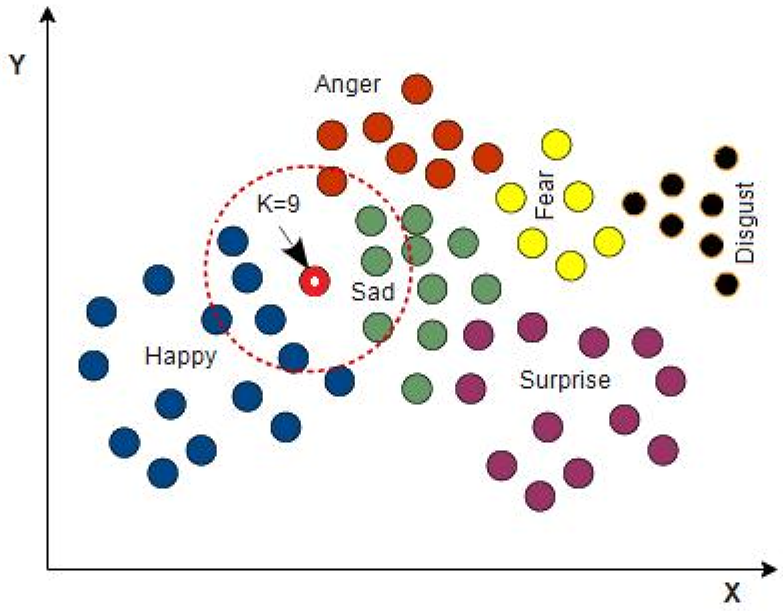

Figure 2: KNN classification with K=9

We aim to use our system to analyze social media and identify impactful emotions trending on Twitter. Following are examples of emotions identified using our system.

Table 5: Emotions Identified from Twitter Stream

No.	TWITTER eWOM	EMOTION
1	Don’t panic. India is banning private cryptos only. Not every crypto 😊!	Happy
2	We unite in grief as this painful tragedy only strengthens our bond and our resolute in fighting terrorism together 👏	Disgust

	We respect, value your sacrifices, we are forever indebted to your sacrifices 🙏	
3	26/11 tells us what hate can demolish and compassion can rebuild. Remembering the victims and saluting the martyrs 💪,	Disgust
4	Market seems crashed! My money is already in loss... ☹	Sad
5	Everyone crying for #cryptoban. Me who never invested in crypto currency 😊	Happy
6	You all know we get a good news from Indian crypto market so, let's celebrate this news with Giveaway 👏	Happy

From the above table, we observe that people talk about various events and express their emotions on social media. AI-based emotion analysis is an alternative to traditional polling and cost-effective solution for decision-makers to understand the situation and respond to any emerging crisis.

4. SENTIMENT ANALYSIS USING NEURAL NETWORKS

High impact opinions often contain emotions or sentiment. Sometimes, explicit emotions may not be present in eWOMs; but, hidden sentiment might be present. Labeling the sentiment of eWOM as positive, negative, or neutral is a classification problem. ‘Sentiment Polarity’ is a context-sensitive meaning in sentiment analysis. Rule-based sentiment analysis systems calculate the sentiment polarity of a message based on the net of positive and negative words expressed about an event but fail to include the context of the event. These systems derive day-to-day sentiment scores by counting positive and negative words. Lexicon based approach involves calculating the sentiment from the semantic orientation of words or phrases in a sentence. For example, the sentence ‘Salute to our warriors who taught us to raise voice against evil’ is a message that contains one positive (salute) and one negative (evil) word. The associated sentiment is not neutral but positive. It is difficult for the rule-based classification system to decide between positive and negative in such a case, since net count of sentimental words is zero. eWOM maybe even worse for parsing because it contains emoticons and special symbols. To alleviate this issue, we use machine learning to predict the polarity and decision-making is done similar to human reasoning.

We present a new architecture for sentiment analysis of messages that operates at the sentence level and uses small convolutions and pooling operations. Supervised learning uses a pre-trained word embedding prepared on a large text corpus. We propose a Polarity Sensitive

Convolutional Neural Network (PSCNN) for eWOM sentiment analysis. In PSCNN, sentence vectors are fed to the convolution and max-pooling layers to generate the document representation. Skip-gram model is best suited for context-based text analysis. The model learns each term within a given context window in a word sequence to capture the skip-gram-based contextual features. Based on the significance learned from the skip-gram model, the PSCNN structures the sentence vectors. As a result, the model applies a non-linear transformation to generate a continuous vector representation for the entire text corpus by extracting high-level semantic information. Convolution in the proposed PSCNN is followed by global max-pooling. We use sequence of word embeddings trained on data collections as inputs to train a CNN-based representation learning model where each sequence of k-words is compacted in the convolutional network. The purpose of our model is to serve the paraphrasing tasks without losing the original context and to match sentences accurately. The system first learns and extracts representations from the two sentences separately. Then it passes the extracted features to max pooling layer to generate a matching degree.

Following are the steps involved in the convolution network.

- The CNN first loads the word embedding as a directory of words to vectors.
- The model then performs a 1-D convolutional operation to learn text representations.
- This representation makes it easy for words with similar meanings to have similar representation.
- The trained model produces a label as output -1 (negative) or 0 (neutral), or 1 (positive).

4.1 CONVOLUTIONS OVER PLAIN TEXT

Let's say, we have a sequence of words $w_{1:n} = w_1, \ldots, w_n$, where each word is associated with an embedding vector of dimension *n*. An one dimansional convolution of width-*k* is the result of moving a sliding-window of size *k* over the sentence, and applying the same convolution filter to each window in the sequence, i.e., a dot-product between the concatenation of the embedding vectors in a given window and a weight vector *u*, which is then often followed by a non-linear activation function *g*.

The concatenated vector of the *i*th window is given by:

$$x_i = [w_i, w_{i+1}, \ldots, w_{i+k}]$$

The convolution filter is applied to each window, resulting in scalar values *r*i:

$$r_i = g(x_i \cdot u) \in R$$

Following is an example of a sentence convolution in the vector-concatenation notation:

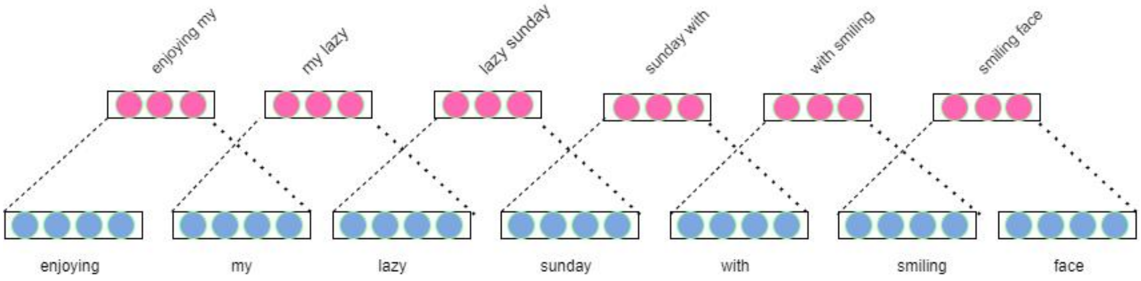

Figure 3: Sentence convolution with *K*=2

The convolution operation produces concatenated vectors which are fed to the neurons to detect similarity.

4.2 EXPERIMENTS AND RESULTS

Our goal is to provide real-life solution using our approach. We aim to evaluate the quality of sentiment classification. We used Twitter data related to product reviews and social events for our experiments. In this section, we describe the evaluation tasks, the data sets used, and the experimental results.

4.2.1 Population and Sample: Our data for evaluation was a set of tweets from May 2021 to July 2021 in the English language originated from India.

Table 6: Sentiment in the Twitter stream

DESCRIPTION	MAY 2021	JUNE 2021	JULY 2021
Total number of tweets	5,423,667	5,212,765	5,314,443
Tweets with sentiment	1,789,810 (33%)	1,563,830 (30%)	1,647,477 (31%)

Tweets with sentiment and hashtag	759,313 (14%)	781,915 (15%)	690,888 (13%)

From the above table, we observe that an appreciable number of tweets with sentiment are available for analysis. Following are examples of sentiment-full messages identified using our PSCNN.

Table 7: Sentiment Identified from Twitter Stream

NO	TWITTER eWOM	SENTIMENT
1	Salute to our warriors who taught us to raise voice against evil 😂!	Positive
2	Their sacrifice for our motherland will continue to inspire the coming generations 👏	Positive
3	She is the hope of Uttar Pradesh 💪, the only leader who is consistently raising the issues and fighting for the people everyday 😊. Young charismatic, with impeccable credentials and history ✌!	Positive
4	The people of UP are urging for relief from gundarj, terror & hypocrisy of Saffron reign ☹	Negative
5	Paranoid, vindictive government will not let farmers survive ☹. Through attacks on farmers the government has finally declared that there is emergency in India now 🙈	Negative
6	Farmers will discuss about APMC mandis in today's Parliament of farmers 😉. Groups of 200 farmers will protest outside the Parliament every day 🙈, during the monsoon session, to strengthen the voice in the temple of Democracy ☹	Negative

5. PREDICTION OF EVENT GROWTH USING MARKOV PROCESS

Presence of emotion or sentiment in twitter stream indicates that a topic is being discussed by public over social media. However, making it useful for decision makers to respond to any emerging crisis depends on the importance of the topic. Is the topic going to escalate as a high impact event and how soon?. How do we measure the momentum?. If we device a method to predict the upward momentum, then such events can be captured for analysis. In this section, we propose an approach to predict the future popularity of the events. Our linear function combines the rate of growth with sentiment to predict the impact. We identify events with upward momentum and present them as high impact events to decision makers.

5.1. Bursty Hashtag Identification

Our approach involves identifying the events using the bursty hashtags which are exhibiting sudden growth within a given time period with consistent sentiment or emotion. We use a tri-state state machine to detect the upward momentum of an event.

S = $\{q_r, q_i, q_d\}$ is a set of states.

q_r is the state corresponding to hashtag growth with sentiment/emotion attached,

q_i is the state corresponding to hashtag growth with neutral tweets (no sentiment /emotion) and

q_d is the state corresponding to hashtag decline (sentiment and emotion may or may not be present).

The tweets are not generated at regular rate in a day. People will be active during the daytime and more tweets will be generated during the day compared to night. We normalize the occurrences of hashtag ht_i for every time window as follows:

n(ht_i) = tw(ht_i) * 100 / T

where tw(ht_i) is the number of tweets containing hashtag ht_i during the window and T is the total tweets generated during the window i.

We model a tri-state state machine as shown below to analyze the hashtag growth/decline.

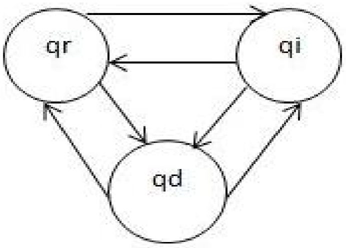

Figure 4: Tri-State Machine for Event Detection

This state machine accepts sequences $\{q_d(q_r|q_i)^* q_d\}$. The event detection is to identify the set of state transitions where n(q_r) > n(q_i). i.e we identify hashtags which have the state transition sequence like $\{q_r q_i q_r q_r q_i q_d\}$ where the emotion/sentiment inversion is minimal. For a given window w_i from Twitter stream, let r_i be the number of tweets containing hashtag *ht*. Then, the probability of transition to state q_r, can be calculated using binomial distribution.

$P(q_r, ht) = \binom{n_i}{r_i} \; p^{r_i} \; (1-p)^{(n_i - r_i)}$

where p is the expected probability of tweets containing hashtag ht with emotion retention. Using Bayes' theorem, we get

$p = P(ht \mid E_r) = P(ht \cap E_r) / P(E_r)$

where E_r denotes the event corresponding to emotion retention (no change in emotion during the interval w_i).

Considering the large volume of tweets published at any time, it is reasonable to approximate this to normal distribution. But, the curve is not perfectly bell- shaped due to external factors affecting an event and sudden fall of tweets (for example, end of football match). It is not possible to predict the growth of hashtag using standard probability models. The random behavior of hashtag growth combined with users' sentiment makes the decision making process a complex activity.

5.2 Markov Process

Our objective is to identify events which remain hot and retain the direction of sentiment polarity at every interval. A hot discussion on cricket match might suddenly stop if the match is canceled due to rain or bad weather. Since the system changes randomly, it is generally impossible to predict with certainty of an event from tweets using standard linear statistical models. After careful analysis of the tweet behavior, we decided to use Markov chain to predict the future. The twitter sentiment resembles a Markov chain called 'drunkard's walk'; a random walk on the number line. At each tweet, the sentiment of public may change from one state to other with equal probability. We model the twitter event burst as Markov chain. Following Markov chain on a countable finite state space represents our stochastic prediction model.

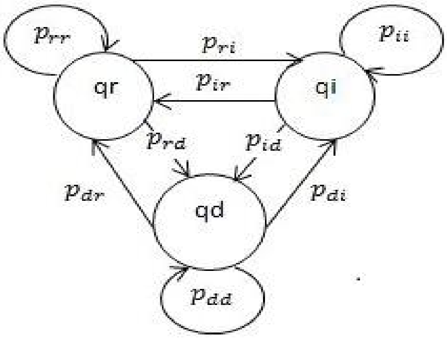

Figure 5. Transition Probabilities of the State Machine

When M is in state q_r, hashtags exhibit growth with sentiment/emotion.

When M is in state q_i, hashtags exhibit growth with neutral tweets.

When M is in state q_d, hashtags are declining.

The transition matrix Mt for stage t of Markov chain is given by

$$Mt = \begin{pmatrix} p_{rr} & p_{ri} & p_{rd} \\ p_{ir} & p_{ii} & p_{id} \\ p_{dr} & p_{di} & p_{dd} \end{pmatrix}$$

The probability for transition from state q_r to q_r denoted by (p_{rr}) is given by the formula

$p_{rr} = \sum_{i=1}^{n}$ Pi(ht | E_r) / n

The probability of a hashtag ht to be burst after n intervals is predicted using Markov chain

$Mt^{(n)}$ = (1 0 0) $Mt^{(n-1)}$

However, along with the Markovian transitions if the full history of the previous transitions is taken into account for prediction, it will provide powerful clues about the likely next stage. We combine learning with Markov transition and develop an Additive Learning Markov chain. In this process, at every interval a deviation of estimated E(p_{rr}) from the actual A(p_{rr}) probability is calculated. We add a correction factor to p_{rr} at every step to smoothen the transition probability.

Et(p_{rr}) = $\sum_{i=1}^{n}$ {Ai(p_{rr}) – EiA(p_{rr})} / (t-1)

The correction factor is the average of deviations between estimated and actual probabilities of state transitions of all previous intervals assuming that we have training data set with history of n intervals. The learned probability will not in general be different from the empirical probability, as the model might choose to converge. As the learning continues, the estimated and actual probabilities converge to a common value when we have sufficiently enough data to train. The steady state transition matrix using the Markov process is used to predict values for future intervals.

5.3 Experiments and Results

Twitter messages are short in size. They spread very fast but short lived. For example, many users would discuss about a football match during the match or within few hours right after the match but not for many days after the match. Our data set for evaluation was set of tweets related to India geographic region. Below chart shows the popular hashtags during a period of 24 hours. Analysis shows that the tweets with sentiment are more than neutral tweets for any hashtag.

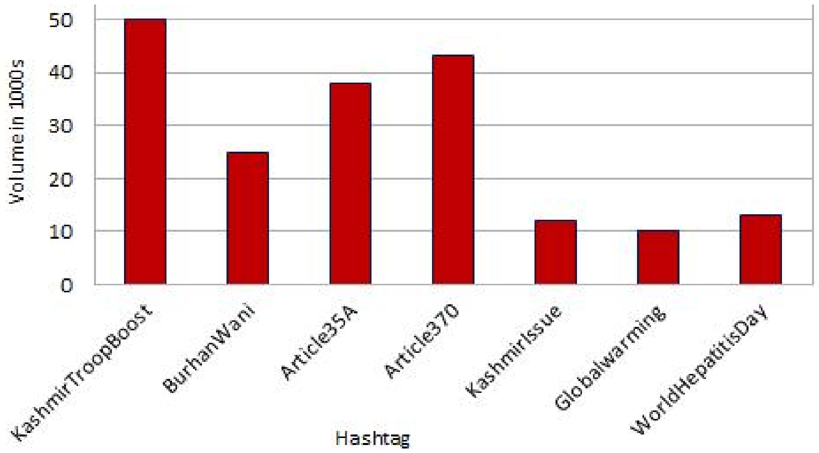

Figure 6. Popular Hashtags in a day

Majority of hashtags having sentiment is an indication that people always express their opinions and emotions in tweets. Opinion based prediction is an effective approach to detect high impact events.

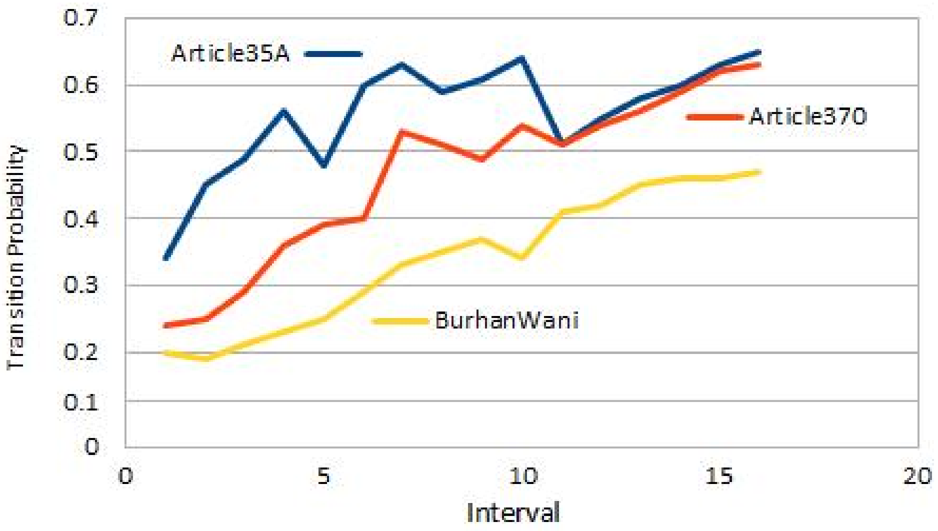

Figure 7: Hashtags with Upward Momentum

After identifying the hashtags which are expected to grow in the next few intervals, we pick up hot events around them. All tweets containing a hot hashtag are combined into a news incident. Following are examples of high impact events indentified using our Opinion Mining System.

Table 8: Events Identified from Twitter Stream

No	HashTag	Sentiment, Emotion	Event	Probability for Growth
1	# BurhanWani	Negative, Anger	1. Pay tribute to the martyrs of Kashmir 2. Marks the martyrdom anniversary 3. Shutdown call by separatists	66%
2	# Kashmir TroopBoost	Negative, Fear	1.Fresh Troops Sparks fear in Kashmir 2. War like situation	71%
3	#Article35A #Article370	Negative, Fear	1.Panic among people 2 Democracy under detention	80%
4	#Kashmir UnderCurfew	Negative, Sad	1. No food, no medicine and no communication 2. Torture every where	73%
5	# Farmers Parliament	Negative, Sad	1. Paranoid, vindictive government will not let farmers survive	61%

After prediction of events, we verified the correctness of prediction by doing a lookup in newspaper passages. From the above table, we observe that people talk about various events and express their opinion in social media. This is really an alternate to traditional polling and cost effective solution for decision makers to understand the situation and respond to any emerging crisis.

6. CONCLUSION

In this paper, we discussed an approach based on opinion to identify high impact events in Twitter. We presented a method to collect, analyze tweets and identify high impact events. As social media being used by people heavily nowadays to express their opinion and emotions, an artificial intelligence based event detection system is definitely a need of the hour. Our Markov based approach identifies events of interest using stochastic process. The result set can further be used to train a model using machine learning. The data set from our Markov prediction combined with machine learning will definitely be a good event prediction system for the digital world. A machine learning system trained with our events from twitter can also be used to identify topics of interest in other social media like Facebook, Instagram etc.

REFERENCES

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. Proc. Conf.Human Language Technology and Empirical Methods in Natural Language Processing, pages 579–586, 2005.

[2] Neviarouskaya, A., Prendinger, H ., Ishizuka, M. Analysis of affect expressed through the evolving language of online communication. In Proc. of the 12th Int'l Conf. on Intelligent user interfaces. January 2007. Pages 278–281

[3] Bradley, M. M, Greenwald, M. K. Petry, M.C. Lang, P. J. Remembering pictures: Pleasure and arousal in memory. Journal of Experimental Psychology: Learning, Memory, & Cognition. Vol. 18: 379–390, 1992.

[4] R. Plutchik. Emotion: Theory, Research and Experience. In Theories of emotion, volume 11, page 399. Academic Press, 1980.

[5] R C Balabantaray, Mudasir Mohammad, and Nibha Sharma. Multi-Class Twitter Emotion Classification: A New Approach. International Journal of Applied Information Systems (IJAIS), 4(1):48–53,(2012).

[6] W Wang, L Chen, K Thirunaraya, AP Sheth: Harnessing twitter big data for automatic emotion identification: IEEE (2012).

[7] G Sanjiv R. Das, Mike Y. Chen, (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53(9):1375-1388.

[8] Walaa Medhat, Ahmed Hassan, Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal (2014) 5, 1093-1113.

[9] Fan Teng-Kai, Chang Chia-Hui. Blogger-centric contextual advertising. Expert Systems with Applications 2011;38:1777-88.

[10] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems; 2013.

[11] Duric Adnan, Song Fei. Feature selection for sentiment analysis based on content and syntax models. Decision Support Systems; 2012;53:704-11.

[12] Kaufmann JM. JMaxAlign: A Maximum Entropy Parallel Sentiment Analysis Tool. In proceedings of COLING'12; Demonstration papers, Mumbai 2012. Pages 277-88.

[13] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decision Support Systems 2011;50:755-68

[14] Gagniuc, Paul A. (2017). Markov Chains: From Theory to Implementation and Experimentation. USA, NJ: John Wiley & Sons. pp. 2–8.ISBN 978-1-111-38755-8

[15] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In WI-IAT, pages 120–123, 2010.

[16] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, Hongjun Lu. Parameter Free Bursty Events Detection in Text Streams. Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

[17] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14 (4):32–43, 1999.

[18] G. K. Zipf. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Hafner Pub. Co, 1949.

[19] J. Weng and B.-S. Lee. Event detection in twitter. In ICWSM, pages 401–408, 2011.

[20] M. Mathioudakis and N. Koudas. Twitter monitor: trend detection over the twitter stream. In SIGMOD, pages 1155–1158, 2010

[21] Alimehmeti, I. (2021). Efficacy and Safety of AZD1222, BNT162b2 and mRNA-1273 vaccines against SARS-CoV-2. Albanian Journal of Trauma And Emergency Surgery, 5(1), 791-796. doi: 10.32391/ajtes.v5i1.178