International Engineering Journal For Research & Development
Vol.4
Issue 5

E-ISSN NO:-2349-0721

Impact factor : 6.03

STUDY OF VARIOUS IMPLEMENTED APPROACHES FOR RUMOUR DETECTION OVER SOCIAL MEDIA PLATFORM

Mrs.Shital.M.Mohod
Dept.of CS
VidyaBharati Mahavidyalaya, Amravati
Amravatishitalmmohod@gmail.com

Dr.SwatiS.Sherekar
PG. Dept. of CSE
S.G.B.A.U. Amravati University Amravati
SS_Sherekar@rediffmail.com

Dr.V.M.Thakare
HOD ,PG. Dept. of CSE
S.G.B.A.U,Amravati university,
vilthakare@yahoo.co.in

Abstract

The rapid growth of World Wide Web has resulted in substantial increase in use of social media.Social interactions can be inferred on the web using the mailing list and home page links. It also represents the social lives of the individuals, collaborations, communities and relationship. Social networking groups are becoming increasingly important due to the volume and activities. The social media data consists of comments, Feedback and reviews.Rumour detection is used for text classification which classifies text into Positive, Negative and Neutral. The increasing use of social media platforms for information and news gathering, its unmediated nature often leads to the emergence and spread of rumours. The aim of the survey is to provide a study on various implemented approaches for rumour detection over social media.

Keywords- Social Media, RumourDetection,SVM,KNN,NB,ME.

1. Introduction

An online social media is a big platform to share thoughts, ideas and knowledge. Now days the day life stress with social platform and end on it as well this will become big media of communication. This media will share the good as well as bad thoughts. People from different geometric locations to talk, share photos, ideas and interests, or make new friends as a virtual community is a website on the internet that serves as an ultimate location for. Online social network increase in security treats and rumors with the rapid increase in popularity. By exploiting user's privacy, identity and confidentiality the intruders and attackers are able to outsmart the security measures by using several techniques [1]. Numerous information is propagating through on-line social network similarly as every positive and negative information. The information posted on social media not always right or not truthful. All people can share information and also give their opinions on that platform is an advantage of social media. The disadvantage of such rapid diffusion of information is that negative information are also spread [2]. Rumour detection is consider as a binary classification task where predefined set of category of binary class as {Rumour, Non-Rumour}.Automatic resolution of rumours is a challenging task that can be broken down into smaller components that make up a pipeline including Rumour detection, Rumour tracking and stance classification leading to the final outcome of determining the veracity of a rumour [3]. A rumour detection system that identifies, in its early stages, postings whose veracity status is uncertain, can be effectively used to warn users that the information in them may turn out to be false.This study include various classification of rumour detection techniques over social media platforms such as machine learning approaches and Lexical based approaches with their comparative analysis.

2. Rumour Detection Approaches:

In Broadly , there exist two types of methods for rumour Detection : Machine Learning methods often rely on supervised Classification approaches, where rumour detection is framed as a binary(i.e. positive or negative).On the other hand ,Lexicon based methods make use of a predefined list of words. Various implemented techniques for rumour detection over social media are discussed with more detail[4].

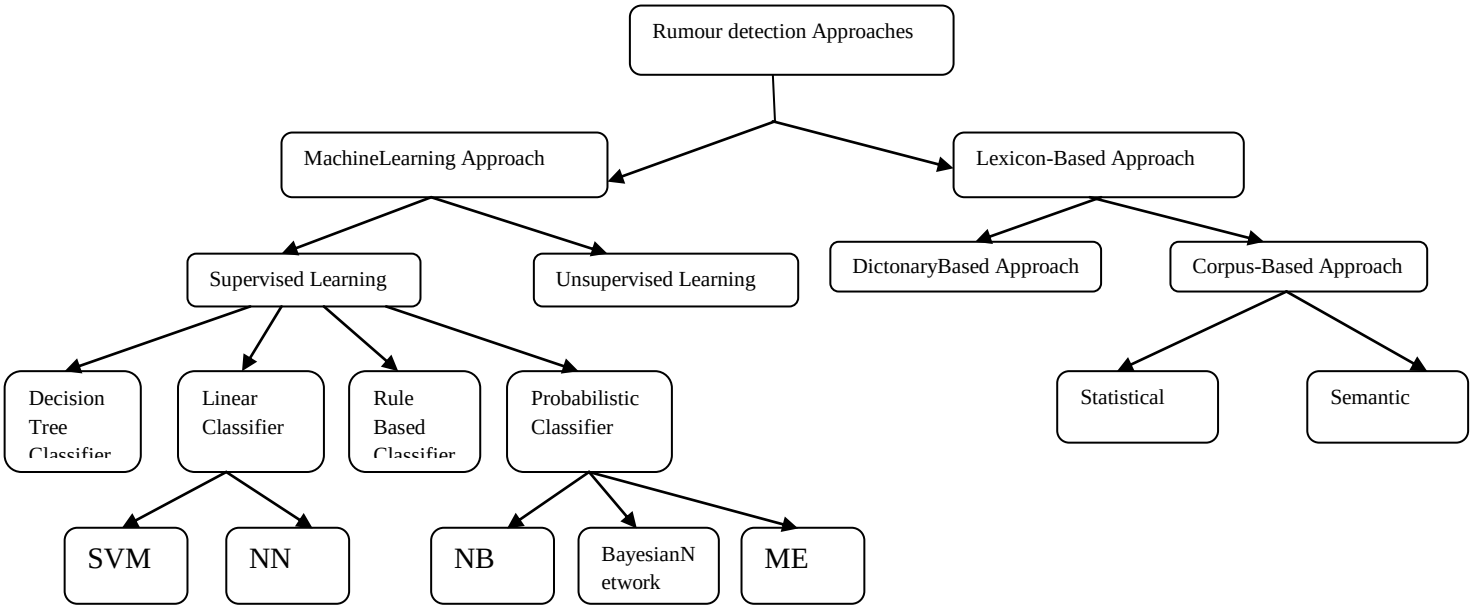

Figure 1: Rumour detection Approaches

2.1. Machine Learning Approaches:

Machine learning algorithms can be addressed as a combination of methods to automatically detect the available pattern in the given set of data. It make use of undiscovered patterns to forecast the future data (or) to implement the decision making under uncertainty. The machine learning techniques are most useful techniques for the rumour detection classification for categorizedtextintopositive,negative or neutral categories,. In machine learning techniques, training and testing datasets are required [5]. A training dataset is used to learn the documents and test data set is used to validate the performance. The content classification strategies utilizing machine learning approach which is typically divide into supervised and unsupervised learning strategies. Supervised learning is performed by considering the target value (i.e. label) and unsupervised learning is conducted by not considering the target value(i.e. label).There are various types of algorithms for supervised learning such as classification(SVM, Decision tree, Naivebayes,ANN etc) and unsupervised learning algorithm such as clustering (MaximunEntropy,Basian Network).

2.1.1 Decision Tree Classifier

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.[6].

2.1.2 Support Vector Machine:

SVM is a discriminative classifier considered as the best text classification method. It is a statistical Classification method proposed by Vapnik. SVM map input feature vectors into higher dimensional feature space through some non linear mapping. SVM have the ability to update the training pattern dynamically whenever there is a new pattern during classification. SVM's can learn a larger set of patterns and able to scale better, because of classification complexity it does not depend on dimensionality of the feature space. SVM have the ability to update the training pattern dynamically whenever there is a new pattern during classification SVM's can learn a larger set of patterns and able to scale better, because of classification complexity it does not depend on dimensionality of the feature space. SVM have the ability to update the training pattern dynamically whenever there is a new pattern during classification [7].The algorithm is based on the structural risk

minimization principle. It is a supervised learning method widely used for classification and regression tasks.SVM is using the concept of train and test dataset [8].The classifier will be trained with target values and features intrain set. The trained classifier will be tested with new features without target value. The algorithm will produce high dimensionofgeneralizationthantheoriginalsetofdata.

2.1.3Artificial Neural Network [ANN]:

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computervision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.ANN is a tool to develop machine learning applications [9]. It is also called as multi-layer perceptions. Input, output, and hidden layers are the part of ANN. The hidden layer will perform operations related to the given problem. The user can have multiple hidden layers to get the optimum results. ANN environment provides Feed forward, Recurrent, Convolutional, Botlmann machine, and Hopfield networks to the users.

2.1.4Naïve Bayes:

Naïve Bayes is a simple and easy but effective classification algorithm. It is mostly used for document level classification. The basic idea is to calculate the probabilities of categories given a test document by using the joint probabilities of words and categories. Naïve Bayes is optimal for certain problem classes with highly dependent features. It uses Bayes theorem to predict the probability which migt be offer feature set belongs to a selected label [9].

2.1.5Bayesion Network

In essence, Bayesian means probabilistic. The specific term exists because there are two approaches to probability. Bayesians think of it as a measure of belief, so that probability is subjective and refers to the future.Frequentists have a different view: they use probability to refer to past events - in this way it's objective and doesn't depend on one's beliefs. The name comes from the method - for example: we tossed a coin 100 times; it came up heads 53 times, so the frequency/probability of heads is 0.53.

2.1.6Maximum Entropy:

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naive Bayesclassifier that we discussed in the previous article, the Max Entropy does not assume that the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy [10]. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

2.2Lexicon Based Approach:

In unsupervised technique, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data.

The basic steps of the lexicon based techniques are outlined below [2]:

1. Preprocess each text.
2. Initialize the total text sentiment score: s ← 0.
3. Tokenize text. For each token, check if it is present in a sentiment dictionary.

(a)If token is present in dictionary,

i. If token is positive, then s ← s + w.

ii. If token is negative, then s ← s - w.

4. Look at total text sentiment score s,

(a) If s > threshold, then classify the text as positive.

(b) If s < threshold, then classify the text as negative.

2.2.1 Dictionary-based Approaches:

In dictionary based techniques the idea is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WorldNet dictionary for their synonyms and antonyms. The newly found words are added to the seed list. The dictionary based approach have a limitation is that it can't find opinion words with domain specific orientations [11].

2.2.2 Corpus-based Approaches:

Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based methods can produce opinion words with relatively high accuracy. Most of these corpus based methods need very large labeled training data. This approach has a major advantage that the dictionary-based approach does not have. It can help find domain specific opinion words and their orientations. The final weight of each individual sentence is calculated after considering the whole sentence structure, contextual information and word sense disambiguation [11].

3. Comparative study

Provide comparative studyof existingtechniques for rumourdetection includingmachine learningapproaches and lexicon-basedapproaches, together with evaluation factors[12, 13]. Comparative analysis of Rumour detection classification techniques on variant parameters i.e Techniques, Classification used, Advantages, Disadvantages.From this comparative study It is also observed that different techniques can be combined to overcome each others limitation and provide a better classification all around. Given Table shows the comparative study of existing techniques for rumourdetection[14,15].

Table 1: Comparative Analysis of Rumour Detection Approaches

Techniques	Learning Methodology	Advantages	Disadvantages
Decision Tree	Supervised	*This is very fast in Learning data. *Easyfor understanding purpose.	* It has problem that it is difficult to handle data with noisy data. * over fitting of data.
SVM	Supervised and unsupervised	*High dimensional input space. *Few irrelevant features. *document vectors are sparse.	* a large amount of training set is required. * Data collection is tedious.
KNN	Supervised	*Based on the fact that the classification of an instance will be somewhat similar to those nearby it in vector space. *it is considered computationally efficient.	*Large storage is required. *Computationally intensive recall.
Naïve Bayes	Supervised	*Simple and intuitive method *It combines efficiency with reasonable accuracy.	*Mainly used when the size of the training set is less. *It assumes conditional independence among the linguistic features.
Maximum Entropy	supervised	*This method do not assume the independent features like NB method. *can handle large amount of data.	*Simplicity is hard.
Dictonary Based	Unsupervised	*In this the newly found words are added to the seed list.	*it can't find opinion words with domain specific orientations.
Corpus-Based	Unsupervised	*It helps to find domain specific opinion words and their orientations.	*This method can produce opinion words with relatively high accuracy

4. Analysis & Discussion

Supervised machine learning techniques have shown relatively better performance than the unsupervised lexicon based methods. However, the unsupervised methods is important too because supervised methods demand large amounts of labeled training data that are very expensive whereas acquisition of unlabelled data is easy. Most domains except movie reviews lack labeled training data in this case unsupervised methods are very useful for developing applications. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms.Supervised text categorization requires the extra effort to predefine the categories and to assign category labels to the documents in the training set. This can be very tedious in a huge

and dynamic text databases. Also, for a supervised categorization, different human experts may disagree when deciding under which category to categorize a given document. This leads us to believe that by nature the ideal multilingual text categorization should be an unsupervised task ratherthan a supervised one. both supervised and unsupervised learning methods have the potentials for multilingual text categorization.

5. Conclusion

Machine learning methods like SVM, NB, Maximum Entropy methods were discussed here in brief, along with some other interesting methods that can improve the analysis process in one or the other way. Rumour detection using syntatic analysis of the text is of great consideration. In the world of Internet majority of people depend on social networking sites to get their valued information, analyzing the reviews from these blogs will yield a better understanding and help in their decision-making. It is also observed that different techniques can be combined to overcome each other's limitation and provide a better classification all around. More work is needed in order to further improve the classification techniques.

.

6. References

[1]Ms.S.M.Mohod, Dr.S.S.Sherekar , Dr.V.M.Thakare , "Review on social security attacks on online social networking for Rumors Blocking " , AIC 2K18 Annual IETE Convention International Journal of Electronics, Communication And Soft Computing Science & Engineering(IJECSCSE) , ISSN 2277-9477, 29 and 30 September- 2018.

[2]. Ms. S.M. Mohod, Dr.S.S.Sherekar ,Dr.V.M.Thakare, "Data Extraction on Social Media With Sentiment Analysis and Classification ",NCETS " Research Journey " International E-Research Journal, ISSN: 2348-7143 , February-2019.

[3].Ms.S.M.Mohod, Dr.S.S.Sherekar ,Dr.V.M.Thakare , "Rumour Verification System using Sequential and Multi-tasking Approaches", Recent Advances in Science and Technology (RAISAT – 2019), 5 and 6 March - 2019.

[4]. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[5].SardarHamidian and Mona T. Diab. 2015. Rumor detection and classification for twitter data. In Proceedings of the 5th International Conference on Social Media Technologies, Communication, and Informatics (SOTICS'15). IARIA, 71–77. SardarHamidian and Mona T. Diab. 2016. Rumor identification and belief investigation on twitter. In Proceedings of NAACL

[6].SeemaChithore, D.A.Phalke,"Asurvey on sentiment analysis Approaches", International Journal of innovative research in Computer and Communication Engineering, ISSN-2320-9801,vol.4,issue 12,December 2016.

[7]. Naïve Bayes MadhoushiZohreh,AbdulRazakHamdan, and SihalilaZainuddin. "Rumour Detection techniwues in recent works. " Science and Information Conference (SAI). 2015. IEEE, 2015.

[8] Abinash Tripathy1 · AbhishekAnand1 ,Santanu Kumar Rath," Document-level sentiment classification using hybrid machine learning approach",KnowlInfSyst DOI 10.1007/s10115-017-1055-z REGULAR PAPER.

[9] ARKAITZ ZUBIAGA, AHMET AKER, KALINA BONTCHEVA, MARIA LIAKATA and ROB PROCTER, "Detection and Resolution of Rumours in Social Media:A Survey," arXiv:1704.00656v3 [cs.CL] 3 Apr 2018.

[10] T. Hashimoto, T. Kuboyama, and Y. Shirota, (2011) "Rumor analysis framework in social media," IEEE Region 10 Conference TENCON 2011, pp. 133-137.

[11] ARKAITZ ZUBIAGA, AHMET AKER, KALINA BONTCHEVA, MARIA LIAKATA and ROB PROCTER, "Detection and Resolution of Rumours in Social Media:A Survey," arXiv:1704.00656v3 [cs.CL] 3 Apr 2018.

[13] ArkaitzZubiaga, Maria Liakata and Rob Procter, "Exploiting Context for Rumour Detection in Social Media," springer, 2017.

[14] Mira Dholariya,Dr.AmitGanatra, Prof. Dhavalbhoi,"A Survey on Sentiment analysis : tools and techniques", International Journal of innovative research in Computer and communication engineering, ISSN 2320-9801, vol. 5, Issue 3, March 2017.

[15] Ms.A.M.Abhirami,Ms.V.Gayathri," A survey on sentiment analysis methods and approach", Eighth International Conference on advanced Computing(ICOAC)", ISSN5090-5888,IEEE 2016.