**Research Questionnaire**

**Note:**

* **Fill up *Table 2* for each paper. (COPY TABLE 2 AND PASTE AT THE END OF THIS FILE FOR NEXT PAPER)**
* ***Green* – Write few or more lines of required in your own words.**
* ***Red* – write down the list of what is required and description about each in the list**

|  |  |
| --- | --- |
| **TABLE 1** | |
| ***Reg. No. & Name*** | 17BCE0403 C BHARATH SAI REDDY |
| ***Team No.*** | **18** |
| ***Paper Title*** | Unsupervised feature learning for spam email ﬁltering |
| ***Citation*** *(APA style)* | |  | | --- | |  | | Diale, M., Celik, T., & Van Der Walt, C. (2019). Unsupervised feature learning for spam email filtering. *Computers & Electrical Engineering*, *74*, 89-104. | | Chicago |  | |

|  |  |
| --- | --- |
| **TABLE 2** | |
| **Problem answered in this paper.**  **(1-2 lines)** | The problem answered in this paper is addressing about the data transformation that is happening, prior to machine learning classifiers. Feature representation that keeps class differentiability with lower level space for detecting spam is being implemented or proposed. |
| **Detailed description about the problem**  **(5-8 lines)** | More number of features will generate negative performance on the learning classifier also Computational time for data processing during the training process will be drastically increased due to presence of more number of features in the data. Preprocessing is one of the major step nowadays in any training processes. Data preprocessing stage consists of feature extraction and feature reduction process in the field of machine learning. It also sums up the speed of computation and also plays a key role in improving the classification accuracy. So, with out preprocessing the negative impacts will cause more wrong results to the project. |
| **Why that problem is chosen in this paper? Scope of the problem and solution** (*Refer Introduction*)  **(5-8 lines)** | Spamming is 0ne of the major inconvenience that is faced by every E-mail user. This will also cause loss of data and valuable information. So there are many Spam detection algorithms placed in E-mail clients but spammer always ensure that spam emails dresses as if they are drafted by ensuring professionals. Spammers are also making sure that using words which are mostly used or known to be present in the spam emails are avoided to the core. As a result more feature representation methods will make classifiers to a poor efficient state with some datasets and extremely good efficient state with some other datasets. |
| **History of the problem.** (*Refer Introduction*)  **(8-10 lines)** | Inconsistency of classification performance makes spam filtering systems more prone to attacks. Major spam detection classification makes use of term count based feature construction techniques. Utilization of term count based feature construction techniques such as Term Frequency-Inverse Document Frequency and Bag of Words system will face various limitations which includes curse of dimensionality ,so uncommon terms are to be thrown away . |
| **List of the related/similar problems** (*Refer Related work*) – Describe each with proposed solutions | |
| **Related problem 1** – Describe  **(3-4 lines)** | One of the approaches for email spam filtering utilizes use of Bag of words methodology, which is one of the famous methods that is being implemented for feature representation. This method crawls for unique occurrences of every single word that is present in the document and also identifying the frequencies . The approach has been used in many previous studies. But there are limitations that have to be answered. |
| **Paper in APA style** | |  | | --- | |  | | Ergin, S., & Isik, S. (2014, June). The assessment of feature selection methods on agglutinative language for spam email detection: A special case for Turkish. In *2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings* (pp. 122-125). IEEE. | | Chicago |  | |
| **Related problem 2** – Describe  **(3-4 lines)** | There are few other algorithms that showed better working of feature representation than Bag of differs but still lag with sparsity and they also don’t capture word order. The Term Frequency-Inverse Document Frequency with Bag of differs because it evaluates word significance in the document. |
| **Paper in APA style** | |  | | --- | |  | | Ferreira, A., & Figueiredo, M. (2011, April). Efficient unsupervised feature selection for sparse data. In 2011 IEEE EUROCON-International Conference on Computer as a Tool(pp. 1-4). IEEE. | | Chicago |  | |
| **Related problem 3** – Describe  **(3-4 lines)** | The disposing of normal terms in the dataset is known as stop have revealed an improvement on the performance of classifiers . They are observed to be non-instructive for recognizing spam messages from non-spam messages, notwithstanding, their expulsion lead to loss of report/express structure. |
| **Paper in APA style** | |  | | --- | |  | | Basavaraju, M., & Prabhakar, D. R. (2010). A novel method of spam mail detection using text based clustering approach. *International Journal of Computer Applications*, *5*(4), 15-25. | |  |  | |
| **Related problem 4 –** Describe  **(3-4 lines)** | There is one of the generally utilized multivariate factual method in the content preparing issues which is related with the  singular value decomposition  is known as the Principal Component Analysis. The main aim of this method is Principal Component Analysis, The main goal of it is to transform the feature representation X into the new space of uncorrelated variables |
| **Paper in APA style** | |  | | --- | |  | | Xie, J., Chen, W., Zhang, D., Zu, S., & Chen, Y. (2017). Application of principal component analysis in weighted stacking of seismic data. *IEEE Geoscience and Remote Sensing Letters*, *14*(8), 1213-1217. | | Chicago |  | |
| **Related problem 5 –** Describe  **(3-4 lines)** | Latent Semantic Analysis , It was acquainted with diminish the quantity of measurements by catching shrouded designs inside documents .The main objective of Latent Semantic Analysis is to find semantic meaning behind words. Latent Semantic Analysis employs  singular value decomposition  approach to compute feature representation. |
| **Paper in APA style** | |  | | --- | |  | | Qian, F., Pathak, A., Hu, Y. C., Mao, Z. M., & Xie, Y. (2010, June). A case for unsupervised-learning-based spam filtering. In *SIGMETRICS* (Vol. 10, pp. 367-368). | |  |  | |
| **What is the proposed solution in this paper for the problem chosen?** (*Refer Proposed work*)  **(5-8 lines)** | This study suggests to use a neural network based approach to generate a more effective feature representation. Which will direct towards better classification working in a lower dimensional feature space. Distrubuted bag of words and distributed memory learning models are used to learn a numerical vector of fixed length of a each email document .These models capture word ordering and semantic meaning from a text document and CS is used and then auto encoder is utilised |
| **Architecture of the proposed solution.** (*Refer proposed work*) **Diagram** |  |
| **Name of the approach as stated by the authors (if not, you try to give a name based on the concepts used)** | Feature refining Using autoencoder and neural networks |
| **List of existing algorithms used by the authors to complete the proposed work.**  **(1-2 lines for each algorithm)** | DBOW(Distributed bag of words)  DM(Distributed Memory)  SVM[support-vector machines]  C4.5  RF [Radio Frequency]  CS[Cosine Similarity]  Autoencoder |
| **List of datasets used.** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | 1.ENRON data  2.IMDB data  3.Trec07 |
| **References/links to each of the dataset used in this paper (in APA style)** | 1. <https://ars.els-cdn.com/content/image/1-s2.0-S0045790617323832-mmc1.zip>  2. <https://ars.els-cdn.com/content/image/1-s2.0-S0045790617323832-mmc1.zip>  3 <https://ars.els-cdn.com/content/image/1-s2.0-S0045790617323832-mmc1.zip> |
| **Why the above dataset(s) used?** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | ENRON data has a total of 33, 716 emails in which 17, 171 are labeled as spam  The IMDB data is made up of 50000 movie reviews.  Trec07 has a very good quality data for the required constraints |
| **List of equations that are very well applied in this problem domain** | Equation 1: cos θ = x(i) · x(j) /||x(i)|| ||x(j)||.  Description: computes the angular distance between two document vectors. This can help in finding similarity.  Equation 2: *J***x** (*i* ) ) 2 ,  Description: Minimization of J is the main task. where omega and b are network parameters.  Equation3: K(x (i ) , x (j) ) = theta(x (i ) ) · theta(x (j) ) .  Description: k(z , x ) = z T x those terms are kernel parameters. |
| **List of method(s)/metrics used to evaluate the proposed approach.** (*Refer experimental evaluation/result discussion*)  **(5-8 lines)** | 1.Precision  It tells us about What proportion of positive identifications was actually correct?  2. Recall  What proportion of actual positives was identified correctly?  3. F – measure  It is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. |
| **List of supporting tools/concepts**  **(3-4 lines)** | SVM  LSA  RF  C4.5  DBOW DM  Auto encoder |
| **What are the similar approaches with which the proposed approach is compared?**  (Refer experimental evaluation/result discussion)  **Explain each of these approach**  **(3-4 lines)** | Approach/method 1: DM + DBOW  Description : F-score % is little less when compared with other approaches and falls down at a certain point.  Approach/method 2: , DM + DBOW and CS + Autoencoder  Description : F-score % is high when compared with other approaches  Approach/method 3: , BoW+IG  Description : F-score % is too low when compared with other approaches |
| **How the results of proposed approach are compared with other similar approaches?** (*Refer experimental evaluation/result discussion*) | They are compared with different data sets and also by calculating the F-score,perscion and recall and also plotting a graph for each data set with all the different approaches with F-score and feature size. The results are also analyzed by plotting between True positive rate and False positive rate for all data sets. |
| **Advantages/merits of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | The proposed method, distributed memory and distributed bag of words with cosine similarity and Autoencoder, had shown to be effective approch baseed on data which used to study these and for performance analysis and better understanding and training of the mehod . This approach is far more efficient than many approaches that are being implemented |
| **Disadvantages/limitations of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | One of the major problem that is being faced is regarding the proposed method is that the cosine similarity measure may lay path to a huge feature space before starting of the Autoencoder phase for feature reduction. This problem may result in space complexity issue which is not good for any application which will display time lag during real time implementation. Embedding and cosine similarity measure issue have to solved. |
| **Future work as stated by authors**  (*Refer conclusion / result discussion / experimental evaluation*) | There is still the problem of word embedding and cosine similarity measure issue.  Clustering algorithms and random sub-sampling approach has to be studied in order to achieve more efficiency and how it will be implemented in the future.Long term and short term memory problem has to be studied further for better performance. |
| **Your one page write-up about the paper** | |
| Email spam, also known as junk email, is unsolicited bulk messages sent through email. The use of spam has been growing in popularity since the early 1990s and is a problem faced by most email users. Recipients of spam often have had their email addresses obtained by spambots, which are automated programs that crawl the internet looking for email addresses. Spammers use spambots to create email distribution lists. A spammer typically sends an email to millions of email addresses, with the expectation that only a small number will respond or interact with the message.  Spam detection is a simple yet a major and still a trouble causing problem in the world, it comes under a classification problem which either classifies the email as a spam or no spam .We know there are a excessive number of methods to identify an email whether it is a Spam or Not ,but here authors suggest a new and efficient way with a less complexity and also by reducing the number of features ,we know excessive number of features will affect the performance of the learning classifier negatively.Computional time also matter as it has to be a real time solution ,as soon as the email comes to inbox it has to classify the email as soon as possible .The authors suggest a new way by using Distributed bag of words(DBOW),Distributed memory(DM),Cosine Similarity, Autoencoder and the algorithms used are SVM,RF,C4.5.  The whole procedure of recognizing spam email from different beneficiaries as proposed in this examination utilizes seven fundamental strategies in particular:  Data Collection and Data preprocessing which is the initial phase  Unsupervised Feature Learning, This is implementing phase  Document Feature representation, It deals with Document feature displaying and tokening  Feature transformation, Deals with transformation of the implemented features  Feature Reduction, Decreasing the Number of features  Hyperparameter Optimization  Classification  The Data preprocessing phase extracts meta-data from the email of the dataset. This preprocessing phase includes expulsion of accentuations and invalid characters, change everything being equal and words to lower case. For the Unsupervised Feature Learning and Document Feature representation Distributed Memory + Distributed Bag Of Words in an unsupervised manner generate and learn continuously distributed vectors that can represent meta-information of the preprocessed email files. Here we are using 3 data sets namely, ENRON,Trec07,IMDB. These data sets are the ones with more efficient classified messages .  But , for Trec07 Cormack data only 70K unlabeled emails are used to train Distributed Memory + Distributed Bag Of Words model.  Autoencoder is then employed to reduce the feature space into more robust feature representation during the Feature Reduction phase. This is to ensure that the information regarding the closeness of email contents is still captured in the lower dimensional feature space.  In this study, the classification algorithms were implemented using Weka which consists of learning algorithms that have the capability to distinguish labeled data. This classification phase included Hyperparameter Optimization which was important for training Support vector machines. RBF kernel function with C = 100 and λ = 0.001 was found to lead to a better performance for Support vector machines. Optimization of hyperparameters was done using grid search algorithm with 10-fold CV.  Performance is calculated with the measures such as Precision , Recall, F-measure, also plots for F-measure % and The feature size is plotted for the better understanding of the working of the three algorithms [Support vector machines,RF,C4.5].Also plots for True positive values and False positive values.Also Distributed Memory + Distributed Bag Of Words and Cosine Similarity +Autoencoder is more efficient and reliable than DBOW+DM. Instead of Support vector machines ,also KNN can be tried. Also Cosine similarity can be replaced with Jaccard similarity. | |
| **Your findings: (possible alternate for the solution proposed)**   * Distributed Memory + Distributed Bag Of Words and Cosine Similarity +Autoencoder is more efficient and reliable than DBOW+DM * Instead of Support vector machines ,also KNN can be tried. * Also Cosine similarity can be replaced with Jaccard similarity .. | |

|  |  |
| --- | --- |
| **TABLE 1** | |
| ***Reg. No. & Name*** | C Bharath Sai Reddy |
| ***Team No.*** | **18** |
| ***Paper Title*** | A new semantic-based feature selection method for spam filtering |
| ***Citation*** *(APA style)* | |  | | --- | |  | | Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. *Applied Soft Computing*, *76*, 89-104. | |

|  |  |
| --- | --- |
| **TABLE 2** | |
| **Problem answered in this paper.**  **(1-2 lines)** | The General Problem faced by every E-mail user, Spam is being addressed and way to avoid it is being proposed in this study. |
| **Detailed description about the problem**  **(5-8 lines)** | Currently, words are being selected by using a method called feature selection , they are then used to create feature vectors for training different Machine Learning approaches. Spammers have found a way to trick this system and also use new words which doesn’t get caught with these classifiers , So this study proposes a new feature selection method that is going to make use of sematic ontology to group or cluster the words into topics for building feature vectors. |
| **Why that problem is chosen in this paper? Scope of the problem and solution** (*Refer Introduction*)  **(5-8 lines)** | In this technology connected world, every human is connected to internet all the time , there are lot of ways people communicate with each other over internet. Starting from Instant messaging ,email, forums and websites and lot more. But these are also misused by some unethical people for delivering disturbing content , advertisements with the help of target ads. Spam takes place in all the platforms but the e-mail spamming became very popular and successful due to its extended use for multiple purposes. |
| **History of the problem.** (*Refer Introduction*)  **(8-10 lines)** | As research and reports show, the percentage of spam emails exceeded approximately 50 percentage of global e-mail deliveries in the first 6 half of 2016. Also few researchers reveled that popular social media networks are being used to distribute spam content and companies are unable to stop them effectively. These diminish the popularity of the internet services .Various spam filtering softwares have been installed by ISP’s such as SpamAssassin,Wirebrush4SPAM and many more. As we all know machine learning approaches commonly consists of Extracting information, discarding confusing, noisy ,irrelevant and inconsistent or redundant data, representing each message as vector of features and using machine learning to classify the messages. Feature selection is one of the powerful techniques in preprocessing stage and has become an essential part of the machine learning due to its capability to identify relevant features and discard irrelevant, redundant, or noisy data. Feature selection methods should also evolve with them to ensure an increase in the overall performance. |
| **List of the related/similar problems** (*Refer Related work*) – Describe each with proposed solutions | |
| **Related problem 1** – Describe  **(3-4 lines)** | Grouping words based on the semantics analysis is never done before and the main goal behind this method is to group word based features into meaning full categories that can be easily used to generate feature vectors |
| **Paper in APA style** | Doerfel, M. L., & Barnett, G. A. (1999). A semantic network analysis of the International Communication Association. *Human communication research*, *25*(4), 589-603. |
| **Related problem 2** – Describe  **(3-4 lines)** | Major disadvantage of avoiding Feature selection is present in the redundancy of features. In an example ‘‘Xanax’’, ‘‘Viagra’’ and ‘‘Cialis’’ would probably would be taken as different independent features. But, these features are mainly dangerous and could damage the performance achieved by some ML classifiers . |
| **Paper in APA style** | |  | | --- | |  | | Hofmann, T. (1999). Probabilistic latent indexing. In *The Proceedings of SIGIR* (Vol. 99, pp. 50-57). | |
| **Related problem 3** – Describe  **(3-4 lines)** | The main goal of using topic features is to unify the representation messages having different terms but the same subject matter. This is one of the major problem or obstacle ahead. |
| **Paper in APA style** | |  | | --- | |  | | Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28. | |
| **Related problem 4 –** Describe  **(3-4 lines)** | Discovering features is one of the main aim for comparing the documents or messages that are found in the data. Choosing a better classifier or model with a suitable method is the key. |
| **Paper in APA style** | |  | | --- | |  | | Nakov, P. (2000, June). Latent semantic analysis of textual data. In *Proceedings of CompSysTech* (pp. 3-5). | |
| **Related problem 5 –** Describe  **(3-4 lines)** | Main problem and disadvantage of previous models is missing of a word to represent the topic in a human readable and understandable form. But, the statistic ways of the topic that is also guessed from samples could be vetted on useless data when the number of files for training is decreased. A completely new method have taken over the utilization of semantic information for aid in the classification of texts. |
| **Paper in APA style** | |  | | --- | |  | | Almeida, T. A., Silva, T. P., Santos, I., & Hidalgo, J. M. G. (2016). Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering. *Knowledge-Based Systems*, *108*, 25-32. | |  |  | |
| **What is the proposed solution in this paper for the problem chosen?** (*Refer Proposed work*)  **(5-8 lines)** | New methodology stated is able to detect and select the features that best summarize the topic of each email. Methodology is divided into four main stages: (i) loading the corpus; (ii) e-mail parsing process; (iii) e-mail topic extractor and guesser; and finally (iv) compute the topic-related significance of each feature. First stage is in charge of loading messages into memory. Parsing is done to remove or extract meta data. E-mail data or message is loaded and analyzed and computes the significance of features. |
| **Architecture of the proposed solution.** (*Refer proposed work*) **Diagram** | This shows workflow of semantic based-feature selection method. |
| **Name of the approach as stated by the authors (if not, you try to give a name based on the concepts used)** | Feature selection using topic extraction methodology  **“Topic guessing”** |
| **List of existing algorithms used by the authors to complete the proposed work.**  **(1-2 lines for each algorithm)** | Naive Bayes  SVM  C4.5  Adaboost C4.5  Bagging C4.5  Random Forests  Logistic Regression  Rough-sets |
| **List of datasets used.** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | 1. CSMINING Spam Emails Datasets  2. Concept drift in e-mail datasets |
| **References/links to each of the dataset used in this paper (in APA style)** | |  | | --- | | 1.CSMINING Group. (2010). Spam Emails Datasets. | | 2. 2. Ruano-Ordas, D., Fdez-Riverola, F., & Méndez, J. R. (2018). Concept drift in e-mail datasets: An empirical study with practical implications. *Information Sciences*, *428*, 120-135. |  | |
| **Why the above dataset(s) used?** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | These data sets are of good quality and are efficient for any classifier algorithms and CSMINING Group. (2010). Spam Emails Datasets is considered by many of the researches and also is used to train many large scale projects as they give more accurate result. |
| **List of equations that are very well applied in this problem domain** | Equation 1:  TM = ⋃ { Ti |∃j ∈ J, hyponym ( si, sj ) ∨ ( si = sj )}  i=1,...,k  Description:    Topic guessing entails grouping message terms into k topics (T1, T2, ..., Tk) where k is the number of syn-sets belonging to any level l ≤ h in the WordNet hierarchy. Each topic Ti is represented by a syn-set si that characterizes its meaning. A topic Ti is present in a message if it contains a term t belonging to a syn-set sj and sj is a hyponym of si or sj = si . Therefore, by using above equation , it is possible to guess the set of topics of a message M containing the syn-sets {sj | j ∈ J} where J is an index set:  Equation 2: Φ = i=1 occurrences(Ti)#TM Description:  where occurrences(Ti) represents the number of times a specific topic (Ti) appears in a message, and #TM depicts the total number of different topics present in the e-mail. |
| **List of method(s)/metrics used to evaluate the proposed approach.** (*Refer experimental evaluation/result discussion*)  **(5-8 lines)** | In order to verify the performance , accuracy and efficiency of feature extraction methodology proposed in this study, they have designed and executed a straightforward and reproducible benchmarking protocol.  They tested various feature selection methods such as Semantic-based Feature Selection, Information Gain, Latent Dirichlet Allocation.  1.Precision  It tells us about What proportion of positive identifications was actually correct?  2. Recall  What proportion of actual positives was identified correctly?  3. F – measure  It is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. |
| **List of supporting tools/concepts**  **(3-4 lines)** | Naive Bayes  SVM  C4.5  Adaboost C4.5  Bagging C4.5  Random Forests  Logistic Regression  Rough-sets  various feature selection methods such as  Semantic-based Feature Selection  Information Gain  Latent Dirichlet Allocation. |
| **What are the similar approaches with which the proposed approach is compared?**  (Refer experimental evaluation/result discussion)  **Explain each of these approach**  **(3-4 lines)** | Information Gain:  Information gain ratio is a ratio of information gain to the intrinsic information. From the results after evaluating , Information Gain is the one of the quickest feature selection method sued in the present world. It is more efficient than other selection techniques. Information Gain was selected due to its widely demonstrated suitability in the spam filtering domain resulting from its good balance between performance, use of computational resources and time consumption  Latent Dirichlet Allocation:  Latent Dirichlet allocation is a new statistical model that allows groups or clusters of observations to be explained by unidentified groups that tells why few features of the data are similar. LATENT DIRICHLET ALLOCATION needs to build a prototype and complex statistical model to successfully obtain the most suitable topics. |
| **How the results of proposed approach are compared with other similar approaches?** (*Refer experimental evaluation/result discussion*) | Information gain being double fast as the Latent Dirichlet Allocation and up to four times faster than the proposed Semantic-based Feature Selection method. The good performance achieved by IG lies in the usage of simple statistical methods to easily compute the importance of each feature .  Finally, the incremented time obtained by method that is stated in the study is boosted by  (i) the rigorous queries to the WordNet ontology  (ii) the use of hyponym/hypernym relations to obtain the respective group or cluster for each word (.  As can be observed, the great changes of time is directly proportional with the complexity of the feature selection method. But , proposed method can be sidelined or used in parallel to reduce the feature selection time and therefore palliate the email-delivery delays. |
| **Advantages/merits of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | 1.Recall values are better for the proposed approach than other compared approaches .  2. The summarized version of recall and precision also validate the satisfactory results of using proposed method, mainly when used in combination with Rough-Sets.  3. Topic guessing methods is also suitable for most configurations using λ values of 1 and 9.  4. Concretely, the combination of proposed model together with Rough-sets achieves the best TCR values having λ values of 1 and 9.  5. Classifiers using this Topic Guessing technique got a better predictive power with respect to Information Gain technique. |
| **Disadvantages/limitations of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | 1. LATENT DIRICHLET ALLOCATION achieves good performance values for Naïve Bayes, SVM, and Logistic Regression classifiers, from a global perspective and better than proposed method.  2.LATENT DIRICHLET ALLOCATION obtains better precision values (avoiding FP errors) in most scenarios  3. IG feature selection method is used are adequate in a real scenario (TCR evaluation < 1). |
| **Future work as stated by authors**  (*Refer conclusion / result discussion / experimental evaluation*) | The proposed classifier achieved a astonishing gain or uptick in performance when the this new approach is implemented . These results stand as evidence that the possibility of identifying spam by using topics instead of words. But, the new feature selection technique is able to remove noise from messages on it own.  Proposed technique helps classifiers to identify more spam messages when using classic approaches. Proposed Topic Guessing technique is able to obtain the astonishing benchmarks when analyzed with other alternatives. This method provides new and more benefits such as perk to know the features ahead of executing any feature selection method.  Manual specified level to establish possible topics is not adequate way of operating and also hierarchical clustering methods will be implemented as stated by the author. Study also suggested it will research more keenly on how obfuscation tricks and classifier poisoning attacks could be efficiently detected during the feature selection stage . Applications of this feature selection method are applicable in many other disciplines. |
| **Your one page write-up about the paper** | |
| A new technique able to identify and select the features that best summarize the topic of each email.The workflow of proposed methodology is divided into four main stages:  (i) loading the corpus  (ii) e-mail parsing process  (iii) e-mail topic extractor and guesser  (iv) compute the topic-related significance of each feature  Initial phase is process of loading data into memory. Once loaded, the header and body parts of each email are extracted from the original message and analyzed .  To minimize the computational overhead and discard nosiy information,  (i) remove the possessive forms of tokens  (ii) throw away non meaning full tokens  (iii) delete tokens included in a list of English proper nouns as well as those containing Uniform Resource Locators (URLs) or digits.  During the 3rd stage, all leftover words are used to identify the topic or topics that best match each message. To finish this task, they have used the WordNet Lexical Database. The hierarchical WordNet database groups words into syn-sets, provides short meaning and usage examples of every word, defines various kinds of semantic relations between syn-sets.  To find e-mail categories , an assumed hierarchical level (h) should be selected in order to semantically clustering features into more understandable categories . Thereby, topic guessing entails grouping or clustering message terms into k topics (T1, T2, ..., Tk) where k is the number of syn-sets belonging to any level l ≤ h in the WordNet hierarchy. Each topic Ti is represented by a syn-set si that characterizes its meaning. A topic Ti is present in a message if it contains a term t belonging to a syn-set sj and sj is a hyponym of si.  So now TM has to be calculated with the help of above parameters.  TM = ⋃ { Ti |∃j ∈ J, hyponym ( si, sj ) ∨ ( si = sj )}  i=1,...,k  After that we find Φ = i=1 occurrences(Ti)#TM With all these parameters and functions the Topic guessing is developed. It has been compared with the other popular selection models such as Information Gain, Latent Dirichlet Allocation. Precision, Accuracy and F-measure have been calculated with different classifier types as Naive Bayes,SVM,C4.5,Adaboost C4.5,Bagging C4.5,Random Forests, Logistic Regression, Rough-sets.All the obtained results are analyzed. The overall version of recall and precision also validate the results of using proposed method, mainly when used in combination with Rough-Sets. Recall values are far better for the proposed approach than other compared approaches . | |
| **Your findings: (possible alternate for the solution proposed)**   * Manual specified level to establish possible topics is not adequate way of operating, automatic selection method has to be developed. * More efficient obfuscation tricks and classifier poisoning attacks could be detected during the feature selection stage * Furthermore, the research of different feature representation schemes with all Machine Lerning techniques are also valuable areas of study | |

|  |  |
| --- | --- |
| **TABLE 1** | |
| ***Reg. No. & Name*** | C Bharath Sai Reddy |
| ***Team No.*** | **18** |
| ***Paper Title*** | Detection of spam-posting accounts on Twitter |
| ***Citation*** *(APA style)* | |  | | --- | |  | | Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. *Neurocomputing*, *315*, 496-511. | |  |  | |

|  |  |
| --- | --- |
| **TABLE 2** | |
| **Problem answered in this paper.**  **(1-2 lines)** | This study analysis has displayed that common spammers and social media bot spammers behave similarly. A realtime spam identification detection method that performs far more efficient than existing methods and models. |
| **Detailed description about the problem**  **(5-8 lines)** | Data is being exploited and put in bad ways by individuals and organizations to gain competitive advantage and financial advantage , a substantial amount of data is being generated by spam or fake users. On an average in every 200 messages there is one Spam message and in every 21 tweets there is one spam tweet. This unimaginable growth of spam messages is compromising research work based on social media data. |
| **Why that problem is chosen in this paper? Scope of the problem and solution** (*Refer Introduction*)  **(5-8 lines)** | Research work is back bone of nay field, a lot of research work is going into vain due to false results obtained from data that contains spam.Also a lot of people are being misleaded by these spam messages and are being manipulated with false news from online. |
| **History of the problem.** (*Refer Introduction*)  **(8-10 lines)** | A report by Nexgate found out that on average one spam post occurs in every 200 social media posts and a more recent research reports that around 15% of active Twitter users are automated bots for spamming . The growing volume of spam content and the use of automated bot accounts to publsih posts raises many problems about the credibility and representativeness of the data for research. Several fields have already identified the crucial role of social media analysis in improving efficiency and gaining competitive advantage. Information extracted from social media has been used in health-care to support effective service delivery, compromising on this data will lead to major problems in a large scale. |
| **List of the related/similar problems** (*Refer Related work*) – Describe each with proposed solutions | |
| **Related problem 1** – Describe  **(3-4 lines)** | Vosoughi observed that both Legitimate and spam news move at same pace. Spam news on Twitter spread swiftly. Social bots are deployed to accelerate the process and human users further amplify the content. |
| **Paper in APA style** | |  | | --- | |  | | Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. | |
| **Related problem 2** – Describe  **(3-4 lines)** | One of the method suggested in a research study uses directed graph modeles to study follower – friend relationships on Twitter and Group more effective spam data. For Big projects They are classified as  Social graph analysis  Text analysis and activity patterns  Analysis of user profile meta-data,  URL usage and the effect of URL obfuscation  Analysis of interaction behavior and URL blacklisting and its effects. |
| **Paper in APA style** | |  | | --- | |  | | Wang, A. H. (2010, July). Don't follow me: Spam detection in twitter. In *2010 international conference on security and cryptography (SECRYPT)* (pp. 1-10). IEEE. | |
| **Related problem 3** – Describe  **(3-4 lines)** | URL’s redirecting to blacklisted sites have grown after an increase in charterers limit .URLs used by spam users and studied how spammers exploit URLs obfuscation to redirect users to malicious sites.Also study have obsereved a huge number of distinct URLs directing to restricted sites due to their involvement such practices. |
| **Paper in APA style** | |  | | --- | |  | | Lee, S., & Kim, J. (2012, February). WarningBird: Detecting Suspicious URLs in Twitter Stream. In *Ndss* (Vol. 12, pp. 1-13). | |
| **Related problem 4 –** Describe  **(3-4 lines)** | Implementation of honeypot accounts that are being controlled by an admin,Some research have created social honeypot accounts masking and behaving as naïve twitter users to entice spam posting users.Who get caught with these are assmed to be spam accounts |
| **Paper in APA style** | |  | | --- | |  | | Lee, K., Eoff, B. D., & Caverlee, J. (2011, July). Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*. | |
| **Related problem 5 –** Describe  **(3-4 lines)** | Identifying spam tweets using the key words, this was one of the old methods that was used. Twitter Spam Drift occurs due to spammers continuously adopt, terminate and update various evasive tricks. Features related to this phenomenon were utilized in training machine learning classifiers. |
| **Paper in APA style** | |  | | --- | |  | | Chen, C., Wen, S., Zhang, J., Xiang, Y., Oliver, J., Alelaiwi, A., & Hassan, M. M. (2017). Investigating the deceptive information in Twitter spam. *Future Generation Computer Systems*, *72*, 319-326. | |  | |
| **What is the proposed solution in this paper for the problem chosen?** (*Refer Proposed work*)  **(5-8 lines)** | As an efficient model ,it should classify test date more efficiently by leveraging the experience obtained from training labelled samples without any mistake. The main goal of the classification is to identify spam users or normal legitimate users more efficiently , by observing their tweets linked with user account meta-data.This study have pivoted hyper parameter values.Correlation is one of the key actor in this technique. |
| **Architecture of the proposed solution.** (*Refer proposed work*) **Diagram** |  |
| **Name of the approach as stated by the authors (if not, you try to give a name based on the concepts used)** | Real time filtering of spam |
| **List of existing algorithms used by the authors to complete the proposed work.**  **(1-2 lines for each algorithm)** | 1.Random forests  2.Exra tress  3.Gradient boosting  4.MaxEnt  5.MLP  6.SVM  7.SVM+MLP features |
| **List of datasets used.** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | 1. Spam-posts detection dataset automated  2. Honeypot  3. Spam-posts detection dataset manual |
| **References/links to each of the dataset used in this paper (in APA style)** | |  | | --- | | All data sets are taken from this link. | | Lee, K., Eoff, B. D., & Caverlee, J. (2011, July). Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*. | |
| **Why the above dataset(s) used?** (*Refer experimental evaluation/result discussion*)  **(3-4 lines)** | Honeypot, the automatically annotated spam-posts detection dataset (SPDautomated) and the manually annotated spam-posts detection dataset (SPDmanual). Honeypot dataset [9] is publicly available and useful for studying spam activity on Twitter. It was utilised both as a dataset per se and for collecting the SPD datasets using keywords. |
| **List of equations that are very well applied in this problem domain** | Equation 1:  T T R = unique tokens in D /tokens in D  Description:  Type-token ratio (TTR) measures the richness of a lexicon in a tweet  Equation 2:  LD = words in D excluding stop words /tokens in D  Description:  Lexical density is a concept in computational linguistics that measures the structure and complexity of the tweet. |
| **List of method(s)/metrics used to evaluate the proposed approach.** (*Refer experimental evaluation/result discussion*)  **(5-8 lines)** | *F-score*  *Precision*  *Recall*  *Accuracy*  he *Receiver Operating Characteristics (ROC) curve* and the *area under the ROC curve (AUC)*. |
| **List of supporting tools/concepts**  **(3-4 lines)** | 1.Random forests  2.Exra tress  3.Gradient boosting  4.MaxEnt  5.MLP  6.SVM  7.SVM+MLP features |
| **What are the similar approaches with which the proposed approach is compared?**  (Refer experimental evaluation/result discussion)  **Explain each of these approach**  **(3-4 lines)** | Here only algorithms are compared apart from methods, All those are compared with different metrics.  In this study data sets are spliced such as different feature sets  SPD-account  SPD-user  SPD-network  SPD-optimized  SPD-all |
| **How the results of proposed approach are compared with other similar approaches?** (*Refer experimental evaluation/result discussion*) | As normally false positive and true positive rates are plotted and analyzed .Visualization of the univariate analysis of correlation of each feature with the target. |
| **Advantages/merits of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | 1.Performance is more efficient and constant across the different classifiers and there is major development over the baseline.  2.Proposed Study can be implemented in any real time application.  3.Study also revealed that spam users tend t be more precise over choosing the way of imitating other users by forming enclave of spammers. |
| **Disadvantages/limitations of proposed solution in your view.** (*Refer conclusion / result discussion / experimental evaluation*) | 1.It can’t determine between legitimate social bot or spamming social bot.  2.If the social bot is fully trained to master all the account details then it is hard to catch them |
| **Future work as stated by authors**  (*Refer conclusion / result discussion / experimental evaluation*) | One of interesting dimension for future work stated by author is to study the effect of the new updated which modifies the tweet length, i.e. increase in the maximum length of tweets [25] on spamming activity. So initially automated spam accounts will face obstacles and problems in generating longer tweets intelligently, thereby making these tweets easier to identify and therefore catching the bot and spamming accounts. These kind of updates always help catching these kind of unethical accounts. |
| **Your one page write-up about the paper** | |
| Platforms, such as Facebook and Twitter, are more powerful in making the internet connectivity. Approximatly 1/3 of the world wide population are estimated to be now connected and within 3 year one-half of the global population will be connected. In 2017 Data price has crossed oil price ,data is the more costly item in the world, who owns the more amount of data is powerful in this world, the company or government with more data can manipulate people’s mind and also make them vote to the person whom they decide. Users of these platforms freely Produce and utilize data leading to huge amounts of data. Several companies have already understood that data is going to be the next revolution and social media analysis is going to play a huge role.  Parallelly the benefits are good but, the rapid growth in social media spam contents always rasies the most disturbing and important questions the quality of research based on analyzing this data. One study suggested that over 15% of the active twitter users are automated bots  This approach use an optimized set of readily available features, independent of historical textual features on Twitter. The employed features are categorized as related to the Twitter account, the user or referring to the pairwise engagement between users. A number of machine learning models have been trained. Recursive feature elimination has been employed in order to ascertain the robustness and the discriminative power of each feature. When compared to past methods this method shows more dominance and efficiency.  Honeypot, the automatically annotated spam-posts detection dataset (SPDautomated) and the manually annotated spam-posts detection dataset (SPDmanual). Honeypot dataset is publicly available and useful for studying spam activity on Twitter. It can be utilized as a dataset and also for collecting the keywords.  Keywords play a major part in diging out specific files from large data of information pool.  Performance is more efficient and constant across the different classifiers and there is major development over the baseline. Proposed Study can be implemented in any real time application. Study also revealed that spam users tend t be more precise over choosing the way of imitating other users by forming enclave of spammers. Also few drawbacks such as it cannot differentiate between legitimate bot and spamming bot.  Also Spam accounts which manually identified are trained to a classifier and then tested ,if the data is fauly then the whole process will go in vain and also login activity can also be considered if it returns any pattern then also it can be classified as spam or bot account. Location history should be taken into play as if the VPN is used by the spammer then the fast changing location will tell us that it is a bot | |
| **Your findings: (possible alternate for the solution proposed)**   * All spam accounts which are manually identified are trained to a classifier and then tested * Also login activity should be taken in more priority. * Location history should be taken into consideration as well as the region of the topic it posted. | |
|  | |