B. TECH. PROJECT REPORT

On

SMS Spam Detection and Classification

BY

Apoorv Kulshrestha

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

November 2017

SMS Spam Detection and Classification

A PROJECT REPORT

Submitted in partial fulfillment of the
requirements for the award of the degrees

of
BACHELOR OF TECHNOLOGY
in

COMPUTER SCIENCE AND ENGINEERING

Submitted by:
Apoorv Kulshrestha

Guided by:
Dr. Neminath Hubballi

INDIAN INSTITUTE OF TECHNOLOGY INDORE
November 2017

Acknowledgement

We acknowledge the cooperation, encouragement and austerity of our project guide, Dr. Neminath Hubballi, whose guidance and knowledge were invaluable for the successful completion of this project. We would like to thank him for the same. However, there are many others who share the reward of this effort simply because it would never have been this good without their help. Having said that, I would like to express my heartfelt gratitude to Nikhil Tripathi and Mayank Swarnkar who have also guided us throughout this project.

Contents

Abstract

With the increase in the SMS traffic, the number of spam messages is also increasing. Spam is also defined differently by everyone according to their taste. This paper aims to classify messages as spam or legitimate, then further categorise spam based on their content, while giving user control over classification. This task is accomplished by two levels of classification viz. primary and secondary. The primary classification is done by a binary classifier whereas a multi-label classifier does the secondary classification. The binary classifier has been experimented with classification models like Logistic Regression and Multinomial Naïve Bayes in combination with vectorisation models like Count Vectorisation and Tf-Idf. The best result for primary classification is achieved in case of Naïve Bayes Classifier in combination with Count Vectorisation. For secondary classification, multiple algorithms have been experimented with. ν-SVM give best results. The whole model has been integrated in an Android application which takes into account the suggestion from the user. The model adjusts according to those user suggestions.

Chapter 1

Introduction

The SMS or text messaging has become very popular these days and is a good alternative to voice calls. Unsurprisingly, this system has attracted a lot of spammers. The number of spam messages grow by 500%[6] every year thus increasing nuisance for the end user.

The most significant change is the widespread interconnection with non-cellular services. From one time password(OTP) to second factor authentication (2FA), these legitimate bulk messages differ slightly from spam messages increasing the complexity of filtering [9].

Now with every business group sending bulk messages, whether a message is a spam or not depends on the end user. For example people who are planning a trip may not consider travel offers as spam and people who go out for eating regularly will be interested in what nearby restaurants are offering and similarly shopaholics will want to know as soon as their favourite store offers discount. Many user perception surveys claim that people do like to receive these offers [11]. Since a binary classification cannot fulfill these needs, a requirement for multi-label classification is inevitable.

Its cumbersome to go through all messages when we just want to look at flight deals or food coupons or discount at one of our favourite shopping brand. It motivated us to build an SMS filtering application that classifies messages into travel, food, online retail etc. so that the user can only look at the offers they are interested in.

The contributions of this paper are :

- For the binary classifier, we experiment with various combination of vectorisers (Count Vectorisation, Tf-Idf) and training models (Logistic Regression, Multinomial Naïve Bayes) and compare the result so obtained.

- We propose a machine learning model with Count Vectorisation and Multinomial Naïve Bayes for binary classification and a model with Count Vectorisation and one-vs-rest ν-SVM for multi-label classification.

- Since collecting an exhaustive and diverse SMS spam dataset is not possible because of privacy implications, we propose a mechanism for the user to report any message as spam of one of the categories or not a spam which updates the classification model.

Chapter 2

Related work

A typical SMS consist of a two major part viz. header and body. The header contains details about the sender and recipient. In the following context, the content of the message is the body of the SMS. Literature related to text message classification can be grouped into Content based filtering and Non-content based filtering. Next two sections Section 2.1 and Section 2.2 describe the Content based filtering and Non-content based filtering respectively.

2.1 Content Based Filtering

Most of the works on spam filtering use content based filtering [11] [7] [1] [4]. It depends upon the content of text like spam words, urls, semantics etc. Yadav et al. [11] proposed a user centric approach that used content based filtering using Bayesian machine learning algorithm with user generated features like blacklisting and white listing, preferred keywords to filter unwanted SMSes and reduced the burden of notifications for a mobile user.

A content based filtering begins with a preprocessor which tokenises the message. The output is passed on to the vectoriser. It uses approaches like count vectorisation, tf-idf etc to convert raw text to vectors. The count vectorizer just maintains the frequencies of the token obtained from the preprocessing phase. Unlike count vectorisation, tf-idf takes into account the importance of the word apart from the numerical statistics. The vector thus obtained is given as input to a machine learning model for classification. The popular techniques used for classification are logistic regression, multinomial naïve bayes(MNB) and support vector machine(SVM).

2.2 Non-Content Based Filtering

Many proposed approaches use a non-content based technique over a content based technique [10] [12]. Warade et al. [10] detected the spam messages by checking mutual relation between the sender and receiver and the content of the messages. If no mutual relation is found between sender and receiver, the message is labelled as spam.

Chapter 3

Spam Detector and Classifier

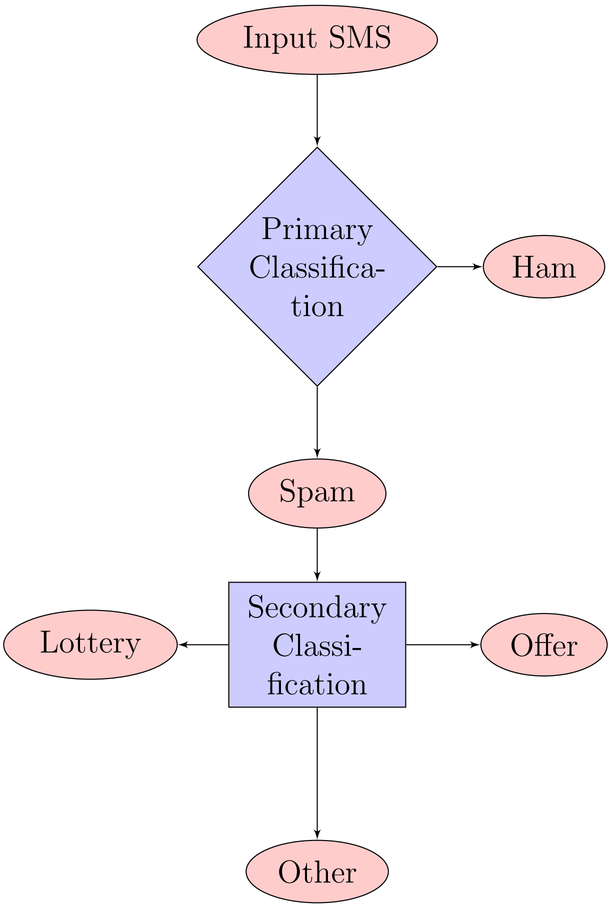


Figure 3.1: Workflow design of the model

In this section we describe the workflow of our project. Here, we propose a model which classifies SMSes based on their content. The entire project model works on two level of classification viz. *primary* and *secondary* classification. The description of these levels are given in Section 3.1 and 3.2.

The complete model is integrated into an android application. On receiving a new SMS, the primary classifier computes it's legitimacy. If the SMS is legitimate, it is displayed in the inbox. Else, the SMS is fed to the secondary classifier for further classification. The application takes into account the suggestion from the user. If the class of the SMS computed by the model does not suit the user, they can suggest changes, and the model adjusts itself to take into account the user preferences.The workflow of the model is shown in Figure 3.1.

3.1 Primary Classification

In primary classification, the input SMS is classified if it is a spam or not. A binary classifier is used as a tool for this classification.

There are many vectorisation and classification techniques that can be used for this classifier. Therefore, we ran an experiment on the dataset we collected and compared the accuracy and AUC score for each of Vectorisation techniques in combination with each of the classification techniques. The combination of Count Vectorisation with Multinomial Naïve Bayes is the one used for binary classifier as it performed better in our dataset. Multinomial Naïve Bayes finds the probability of a SMS d being in class c (*spam* or *ham*) as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

where $P(t_k|c)$ is the conditional probability of token t_k occurring in a SMS of class c and n_d is the number of unique tokens in d.

3.2 Secondary Classification

Some papers like [11] proposed a user centric detection by giving a feature to blacklist or whitelist a message. This was the total extent of user involvement. To the best of our knowledge no previous work proposed a further layer of classification of spam based on the content so that user has more choices and has the message exactly where he wants.

There are many classification techniques that can be used for this classifier. Therefore, we ran an experiment on the dataset we collected and compared the accuracy score for each of the classification techniques. The ν-SVM is used for multi-label classifier beacause it performed better than other models like DecisionTree, ExtraTree etc. SVMs are a set of supervised learning methods used for classification, regression and outliers detection. The advantage of SVM is their abiltiy to learn independent of the dimensionality of the feature space.

Chapter 4

Experiments and Discussion

In this section we explain the experiments performed to evaluate the classification accuracy of proposed approach. The dataset we used in our project is described in Section 4.1. The experiments we conducted for the primary and secondary classification are described in Section 4.2 and 4.3 respectively.

4.1 Dataset

Category	Number of messages
Total	5574
Spam	4827
Ham	747

Table 4.1: Distribution of dataset for primary classification

Category	Number of messages
Spam	747
Lottery	226
Offer	122
Other	399

Table 4.2: Distribution of dataset for secondary classification

We used a dataset by Universidade Estadual de Campinas(UEC) which contains 5574 messages. The dataset is available at `dcomp.sor.ufscar.br/talmeida/smsspamcollection` [3] [5] [2]. It contains SMSes in English language and was pre-labelled as *spam* or *ham*. Of which 747 were spam and the rest were ham as shown in Table 4.1. The secondary classification was something new to be implemented in this project. Therefore, we manually labelled the dataset as *offer*, *lottery* or *others*. The resultant data set after this labelling consists of 122 offer messages, 226 lottery messages and 399 other kind of spam messages . The distribution for the secondary classification is shown in Table 4.2.

Bradley Reaves [9] pointed out that interconnection of sms with non cellular services and the presence of legitimate bulk messages has made filtering even more difficult. They are releasing the largest public dataset to help tackle this problem. Due to unavailability of the above dataset, we are using the pre-existing dataset from other sources.

4.2 Primary Classification

Test Case	Vectorisation Model	Classification Model	Accuracy Score(Percent)		AUC Score(Percent)[8]	
			Average	Best	Average	Best
1	Count	Logistic Regression	97.7107	98.8300	91.2779	95.4100
2	Count	Multinomial Naïve Bayes	98.2699	99.3040	95.3484	99.3241
3	TfIdf	Logisctic Regression	96.0734	97.0611	85.0428	87.9431
4	TfIdf	Multinomial Naïve Bayes	97.6218	98.2211	90.6817	92.5324

Table 4.3: Comparision of Primary Classification Models

	Average	Best
Accuracy Score (in %)	98.2699	99.3040
AUC Score (in %)	95.3484	99.3241

Table 4.4: Accuracy and AUC Score for the binary classifier

For primary classification, as described above, we have two different classification models and two different vectorisation models. To find out the most optimal way, we ran an experiment on all the four combinations of vectorisation and classification models. To evaluate each method, we used **K-fold Cross Validation** technique. In our case, the original sample was randomly partitioned into 4 subsamples. 1 out of 4 subsamples is retained as the validation data for testing the model, and the remaining 3 subsamples are used as the training data. The cross-validation process is then repeated 4 times, with each of the 4 subsamples used exactly once as the validation data. The

4 results thus obtained from each of the fold is then averaged to produce the average accuracy and AUC score for each combination. The best accuracy and AUC score is the one among the 4 folds which scored the best for each combination. As presented in Table 4.3 best classification is achieved when we use **Naïve Bayes Classifier** along with **Count Vectorisation**. Table 4.3 compares each of the combination of vectorisation and classification method based on accuracy and AUC (best and average).The best accuracy and AUC score of our binary classifier obtained are 99.3040% and 99.3241% respectively whereas average accuracy and AUC score obtained are 98.2699% and 95.3484% respectively as shown in Table 4.4.

4.3 Secondary Classification

For secondary classification, as described above, we have multiple inherently multi-label classification algorithms, and multiple binary classification algorithms that can be utilized for multinomial classification through strategies such as *one-vs-rest* and *one-vs-one*. To find out the most optimal way we considered various classification models such as DecisionTree, Bernoulli Naïve, ExtraTree, Gaussian Process Classifier, ν-SVM, Perceptron and Passive Aggressive Classifier. To evaluate each method, we used **K-fold Cross Validation** as described in Section 4.2. The best and average accuracy score for each model is shown in Table 4.5. We can clearly see that nu-SVM outperforms all other models which were experimented. The result worth mentioning about the experiment was that Gaussian Process Classifier both OVO and OVA performed poorly in terms of memory and execution time. Therefore, we used count vectorizer as the vectoriser method along with nu-SVM with linear kernel for the secondary classification. The confusion matrix of ν-SVM is shown in Table 4.7. The best and average accuracy score of our multi-label classifier is 94.30% and 90.08% respectively as shown in Table 4.6. It was created by randomly choosing 123 spam messages for validation and remaining spams were used for the training purpose. We can notice from the confusion matrix that 116 out 123 messages were classified correctly.

Method	Accuracy Score(Percent)	
	Average	Best
DecisionTree	81.71	84.49
Bernoulli Naïve Bayes	83.95	86.63
ExtraTree	80.42	86.63
OVO Gaussian Process Classifier	80.85	82.35
OVA Gaussian Process Classifier	79.46	89.28
OVA ν-SVM	90.08	94.30
Perceptron	81.60	83.95
Passive Aggressive Classifier	78.40	83.95

Table 4.5: Comparison of Multi-label classification models

	Accuracy Score (in %)
Average	90.08
Best	94.30

Table 4.6: Accuracy and AUC Score for the binary classifier

Predicted Actual	Offer	Lottery	Other
Offer	27	2	2
Lottery	2	62	0
Other	1	0	27

Table 4.7: Confusion matrix for ν-SVM

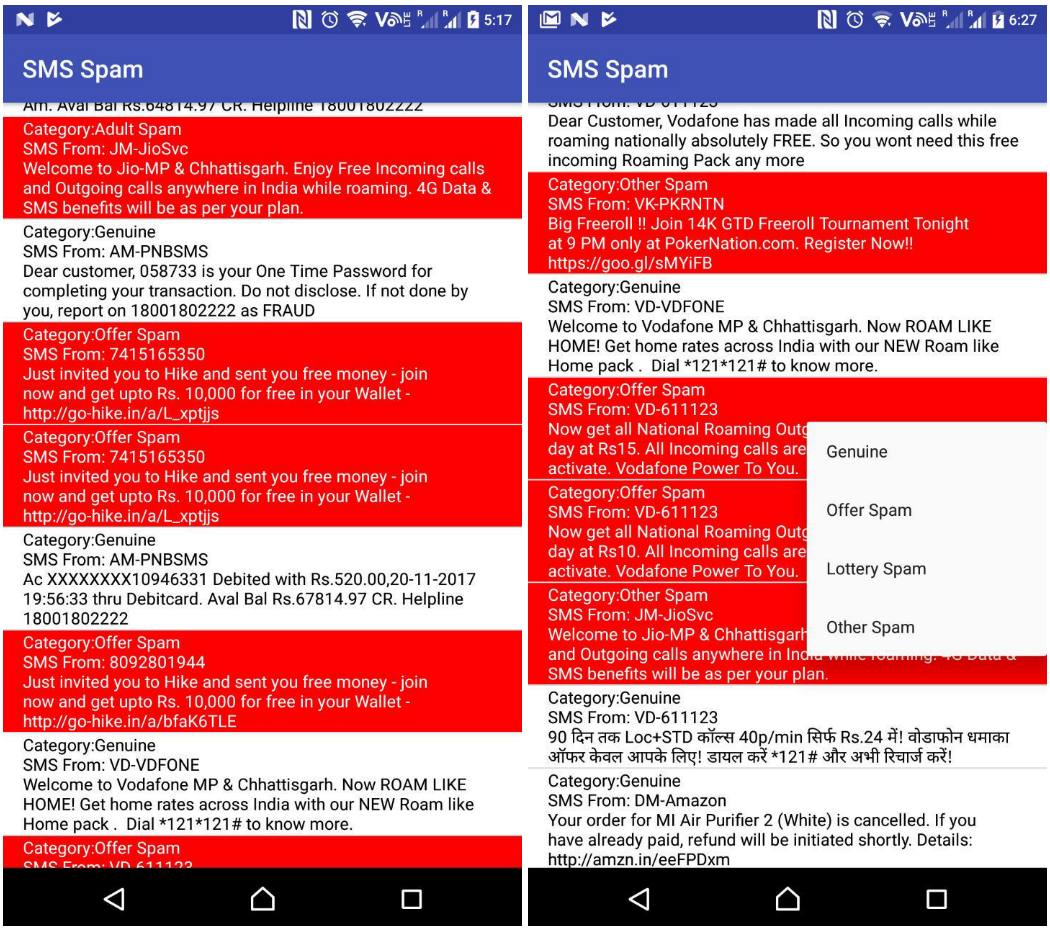
Chapter 5

Android Application

In order to provide all the functionalities described above to the user in a consumable way, an Android Application is developed. Currently, the application is able to fetch existing SMSes and intercept new SMSes, list them for the user as a scrollable list and colour code the SMSes as shown in Figure 5.1. The application also provides the feature for the user to report an incorrectly classified message as spam of a particular category, or as a legitimate message, thereby customizing it for them.

Whenever a new SMS is received, the application intercepts the SMS, and then it does a network call to a server which then connect to a pre-trained model running on the server-side. The SMS is then classified and category where it lies is then returned back to the android phone. Whenever any user tries to change the class of a SMS predicted by the model. Once again a network call occurs, the SMS is sent to the model for re-training. For re-training this SMS is appended to previous dataset. This change persists for future training that is to be done by the model.

Future plans for the application include implementing the training and classification model for a memory constrained environment of a smartphone. Once the user manually tries to change the class of a SMS, the model should not re-train using the whole dataset. It should run the classification with the changed SMS only.



Am. Aval Bal Rs.64814.97 CR. Helpline 18001802222

Category:Adult Spam
SMS From: JM-JioSvc
Welcome to Jio-MP & Chhattisgarh. Enjoy Free Incoming calls and Outgoing calls anywhere in India while roaming. 4G Data & SMS benefits will be as per your plan.

Category:Genuine
SMS From: AM-PNBSMS
Dear customer, 058733 is your One Time Password for completing your transaction. Do not disclose. If not done by you, report on 18001802222 as FRAUD

Category:Offer Spam
SMS From: 7415165350
Just invited you to Hike and sent you free money - join now and get upto Rs. 10,000 for free in your Wallet - http://go-hike.in/a/L_xptjjs

Category:Offer Spam
SMS From: 7415165350
Just invited you to Hike and sent you free money - join now and get upto Rs. 10,000 for free in your Wallet - http://go-hike.in/a/L_xptjjs

Category:Genuine
SMS From: AM-PNBSMS
Ac XXXXXXXX10946331 Debited with Rs.520.00,20-11-2017 19:56:33 thru Debitcard. Aval Bal Rs.67814.97 CR. Helpline 18001802222

Category:Offer Spam
SMS From: 8092801944
Just invited you to Hike and sent you free money - join now and get upto Rs. 10,000 for free in your Wallet - http://go-hike.in/a/bfaK6TLE

Category:Genuine
SMS From: VD-VDFONE
Welcome to Vodafone MP & Chhattisgarh. Now ROAM LIKE HOME! Get home rates across India with our NEW Roam like Home pack . Dial *121*121# to know more.

Category:Offer Spam
SMS From: VD-611123

SMS From: VD-611123
Dear Customer, Vodafone has made all Incoming calls while roaming nationally absolutely FREE. So you wont need this free incoming Roaming Pack any more

Category:Other Spam
SMS From: VK-PKRNTN
Big Freeroll !! Join 14K GTD Freeroll Tournament Tonight at 9 PM only at PokerNation.com. Register Now!! https://goo.gl/sMYiFB

Category:Genuine
SMS From: VD-VDFONE
Welcome to Vodafone MP & Chhattisgarh. Now ROAM LIKE HOME! Get home rates across India with our NEW Roam like Home pack . Dial *121*121# to know more.

Category:Offer Spam
SMS From: VD-611123
Now get all National Roaming Outg
day at Rs15. All Incoming calls are
activate. Vodafone Power To You.

Category:Offer Spam
SMS From: VD-611123
Now get all National Roaming Outg
day at Rs10. All Incoming calls are
activate. Vodafone Power To You.

Category:Other Spam
SMS From: JM-JioSvc
Welcome to Jio-MP & Chhattisgarh
and Outgoing calls anywhere in India while roaming. 4G Data &
SMS benefits will be as per your plan.

Category:Genuine
SMS From: VD-611123
90 दिन तक Loc+STD कॉल्स 40p/min सिर्फ Rs.24 में! वोडाफोन धमाका ऑफर केवल आपके लिए! डायल करें *121# और अभी रिचार्ज करें!

Category:Genuine
SMS From: DM-Amazon
Your order for MI Air Purifier 2 (White) is cancelled. If you have already paid, refund will be initiated shortly. Details: http://amzn.in/eeFPDxm

Figure 5.1: Screenshot of Application

Bibliography

[1] Ahmed, I., Guan, D., and Choong Chung, T. Sms classification based on nave bayes classifier and apriori algorithm frequent itemset. 183–187.

[2] Almeida, T., Hidalgo, J. M. G., and Silva, T. P. Towards sms spam filtering: Results under a new dataset. *International Journal Of Information Security Science(IJISS), 2(1), 1-18* (2013).

[3] Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A. Contributions to the study of sms spam filtering: New collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering* (New York, NY, USA, 2011), DocEng '11, ACM, pp. 259–262.

[4] Gómez Hidalgo, J. M., Bringas, G. C., Sánz, E. P., and García, F. C. Content based sms spam filtering. In *Proceedings of the 2006 ACM Symposium on Document Engineering* (New York, NY, USA, 2006), DocEng '06, ACM, pp. 107–114.

[5] Hidalgo, J. M. G., Almeida, T. A., and Yamakami, A. On the validity of a new sms spam collection. In *2012 11th International Conference on Machine Learning and Applications* (Dec 2012), vol. 2, pp. 240–245.

[6] Murynets, I., and Piqueras Jover, R. Crime scene investigation: Sms spam data analysis. In *Proceedings of the 2012 Internet Measurement Conference* (New York, NY, USA, 2012), IMC '12, ACM, pp. 441–452.

[7] Narayan, A., and Saxena, P. The curse of 140 characters: Evaluating the efficacy of sms spam detection on android. In *Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones &#38; Mobile Devices* (New York, NY, USA, 2013), SPSM '13, ACM, pp. 33–42.

[8] Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[9] Reaves, B., Blue, L., Tian, D., Traynor, P., and Butler, K. R. Detecting sms spam in the age of legitimate bulk messaging. In *Proceedings of the 9th ACM Conference on Security &#38; Privacy in Wireless and Mobile Networks* (New York, NY, USA, 2016), WiSec '16, ACM, pp. 165–170.

[10] Warade, S. J., Tijare, P. A., and Sawalkar, S. N. An approach for sms spam detection. *Int. J. Res. Advent Technol 2*, 12 (2014), 8–11.

[11] Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., and Naik, V. Smsassassin: Crowdsourcing driven mobile-based system for sms spam filtering. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications* (2011), ACM, pp. 1–6.

[12] Zhong, J., Du, J., Yang, Q., Xiang, E. W., and Xu, Q. Sms spam detection using noncontent features. *IEEE Intelligent Systems 27* (2012), 44–51.

Appendix

1. **Tokenization:** Text is divided into tokens, for example by using punctuations and white spaces as token separators.A unique integer ID is given to each token. The vector of all the token frequencies for a given document is used as feature vector to be fed into algorithm.

2. **Tf-Idf:** It is the abbreviation of **Term Frequency-Inverse Document Frequency**. This is a numerical statistic that is intented to reflect how important a word is to a document. **Term Frequency** is the normalised frequency of a word in a document. The frequency can be normalised by dividing by document length or using logarithmically scaled frequencies. The **inverse document frequency** is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. Then tf-idf is calculated as product of term frequency and inverse document frequency.