Text-based Intelligent Content Filtering on Social Platforms

Sadaf Khurshid[1], Sharifullah Khan[1], Shariq Bashir[2]
{12msitskhurshid, Sharifullah.khan}@seecs.edu.pk, shariq.bashir@jinnah.edu.pk
[1] School of Electrical Engineering and Computer Science (SEECS)
National University of Science and Technology (NUST), Islamabad, Pakistan
[2]Department of Computer Science, Mohammad Ali Jinnah University, Islamabad, Pakistan

***Abstract*—Social platforms have become one of the popular mediums of information sharing and communication over the Internet today. People share all types of contents such as text, images, audio and video using these social platforms. Though information gained using these social platforms can be very useful for people around the globe, some of the user generated contents are very negative as they contain abusive, racial, offensive and insulting material. Thus, there is a need for an effective online content filtering technique which blocks these negative contents while not disturbing the access of users to rest of the contents available on these sites. Current techniques simply filter on the basis of URLs blocking and keyword matching or either rely on a large database of pre-classified web addresses. The problem is how to intelligently filter the negative contents, rather than filtering entire websites using their URLs or applying simple keyword matching techniques. In this paper we review a number of existing approaches to content filtering and propose an intelligent content filtering technique that uses sentiment analysis of the text and feature engineering methods to perform text classification.**

***Index Terms*—Content filtering, Feature Engineering, Sentiment Analysis, Social Platforms, Text Classification**

I. INTRODUCTION

In the past few years, social platforms have become a popular interactive medium to communicate, share and disseminate a considerable amount of information [3], [4]. Daily and continuous communication implies the exchange of several types of contents such as text, image, audio and video data [8]. The working rule of these social platforms has become such that any individual who associate with the social platforms freely shares with the world any kind of data considered suitable. ScanSafe's monthly "Global Threat Report" [5] found that up to 80% of blogs on social platforms contained offensive contents and 74% included porn in the form of image, video, or offensive languages. These social platforms have faced continued criticism for publishing a large amount of negative contents in the form of hurtful/abusive/racial/insulting remarks. In order to maintain the Internet a decent place of information it is very important to control the negative contents through some mechanical way as the quantity of data is often tremendous to be effectively controlled by a manual process [17]. Information filtering techniques are commonly intended to deal with substantial volumes of dynamically produced data and present the client with sources of data or information which is likely to fulfill his or her information necessity [17]. Our focus in this research on the information sources such as, reports, remarks and comments.

We have studied the existing filtering techniques in detail and found that most of them use techniques such as filtering of URLs, where it is quite costly to maintain the up-to-date list of new sites. This approach can be easily deceived by using URLs or domain names of the websites which are not relevant or do not depict the type of contents present on the websites [14]. Moreover, blocking through the use of keyword matching also has drawbacks. The issue with the keyword matching systems is that the implications of words rely upon the connection or context of the information. For example, websites about breast malignancy examination or cancer research could be blocked due to the event of the saying "breast" that is utilized as a keyword for "obscenity class" or "pornography category". Not only that, Keyword matching is inclined to spelling mistakes that could be utilized to side step the effective filtering assurance. Lastly rating based filtering techniques are also very common but lack reliability and accuracy as one needs to track whether the content is being rated by reliable, different and an enough number of people rather than certain people rating over and over again [7].

In this paper we have presented a filtering technique, which can overcome the content filtering issues mentioned above. Instead of filtering the contents on the basis of URL blocking, keyword matching or ratings acquired by the users, we have utilized the sentiments of the text and engineered some features in the text to classify the content as negative or abusive. The technique was tested on different data-sets acquired from on-line social platforms and results of our experiments turned out to be quite satisfactory. Our filtering technique is very much suitable for filtering contents on social platforms, such as You-Tube, Facebook and Twitter. They mostly contain lots of text such as reviews and comments which can be useful to mine opinion and determine the acceptability of the contents present over the web [21].

This paper is further organized as follows. Section II presents related work. Our proposed filtering methodology has been

978-1-4799-7505-1/14 $31.00 © 2014 IEEE
DOI 10.1109/FIT.2014.51

explained in section III. To demonstrate the carried out work, our experiments and evaluation results have been explained in section IV. Lastly, we conclude the paper and briefly present some future challenges.

II. RELATED WORK

There have been many efforts done previously in creating information filtering techniques that can be related to the research challenge addressed in our study. They can be broadly divided into three classes:-

A. Keyword Matching Based Filtering Techniques

In [7], a filtering technique was presented which is based on text classification, where forbidden web pages are used as a sample to classify web pages that are required to block. The documents or web pages are represented in vectors based on frequency of the words and then classified as blocked or unblocked by calculating the cosine similarity between the two documents. The authors in [14] have implemented a content based web filtering system which used pornographic web pages as a case study and combined the knowledge gained from their analysis with the working of artificial neural network model to filter the web pages. Neural networks model was used to create a knowledge base of the filtering system, which was trained by taking the samples of both pornographic and non-pornographic web pages. The development of a maintainable filtering system has been presented in [6], which is based on the approach of the expert systems. It easily maintains a knowledge base of that filtering system without the help of knowledge engineers. When the knowledge in the system is failed, the new knowledge is acquired which does not require lots of training.

The above techniques are using similarity value calculations and keyword matching techniques to filter out the documents. Filtering on the basis of a single similarity value cannot achieve very accurate results. Secondly when considering the expert system approach, though the technique gives very accurate results as it is domain specific. However, the maintenance of an up-to-date knowledge base can be a huge problem when applying the same technique on social platforms. Moreover, the over all filtering techniques apply matching of keywords between the knowledge base and the content to be filtered. Thus, to filter content without knowing the semantics of the text can give highly inaccurate results because the use of the words rely upon the connection or context of the information.

B. URL Based Filtering Techniques

In [26] the authors have proposed a URL based filtering technique, in which a compression algorithm is used to compress each URL so that the memory requirement to save the URLs blacklists can be reduced. Moreover, they have also used a prefix or multiple string matching algorithm which enhances the URL look-up performance by matching the prefixes. The major drawback in their technique is that if a URL containing a domain name is in the blacklist, then all the paths and sub-paths of this URL are blocked. For example, if URL www.youtube.com/watch?v=X is listed in blacklist then through this technique, its other paths such as www.youtube.com will also be blocked.

A combination of content filtering and URL filtering technique was used in [19], first of all the incoming URLs are blocked through comparison with already maintained blacklist of URLs, and later on content filtering approach is applied in which the content of web page is filtered by matching the words with an already maintained data structure of sensitive words. The system basically uses combination of both keyword matching and URL filtering approach without taking into consideration the semantics of the text available on the sites. Since, the URL of the website is first matched with the blacklist of URLs and then the keywords are searched inside the content of the website to determine its acceptability of the site; thus, this approach can be very expensive in terms of time and cannot be suitable to apply on social platforms.

C. Recommender System Based Filtering Techniques

In [18] the authors have proposed a video recommender system which allows users to share and uploads videos. The system uses supervised classification algorithms to filter spammers and promoters on the basis of attributes extracted from videos and user profiles. In another research [12] a user rating along with sentiment analysis technique have been applied by using a set of predetermined polarity terms to recommend educational content in social environments. The technique makes the result more accurate as the text is analyzed and polarity is also judged through sentiments before recommending the content to another user. Recommender systems mostly utilize user ratings to filter the content, which might not be very reliable source of filtering as there is always a chance of false ratings generation due to the presence of promoters of the content.

III. PROPOSED METHODOLOGY

However, the techniques of supervised classification and sentiment analysis can be very helpful in filtering content on social platforms. The aim of this work is to propose a novel content filtering approach which combines the techniques from different areas of study. The approach uses techniques from opinion mining, sentiment analysis, feature engineering and classification of text through machine learning algorithms. The combination of these techniques help determine the sentiments of the text in more reliable way; as the sentiments are found not only through user opinions, but also through the use of lexical resource and feature calculations. Our proposed methodology consists of 4 major phases. The first phase is of pre-processing the textual documents and to make them ready to use for sentiment analysis. The second phase is to calculate sentiment scores. Once the polarity scores have been determined, certain features are engineered to help classify the text. Lastly, different machine learning algorithms are used to classify the text into positive/negative class.

A. Document Pre-processing

Distinctive pre-processing methods have been employed to expel the noise from the information set. They serve to reduce the size of the information set and consequently help to speed up the filtering process.

POS tagging: This initial step in text processing is a part of speech tagging in which individual words are mapped to its associated lexical classes, for example, noun, adjective, verb, adverb etc. We labeled each word in the information set with its grammatical feature or parts of speech utilizing Stanford POS Tagger [23].

Tokenization: Tokenization is the process of breaking record sentences into pieces, called tokens. At the same time tokenization helps to discard certain characters which will not be useful further, for example, punctuations. This step identifies the basic textual units which need not be further decomposed [25].

Stopword Removal: Stopwords are also known as noise words, which are of very little significance in setting of categorization and sentiment analysis. For example; articles, conjunctions such as *the*, *is*, *at* and *on*. Since these words are common and having their presence in the data would enormously expand the span of the list without enhancing accuracy or precision [13], thus they have been removed from the data.

There is no definite arrangement of stopwords utilized for natural language processing. We expelled all the stopwords from our data-set utilizing the list from Rainbow (a program that performs statistical text classification) [16].

Bag of Words (Term Frequency): Each document in the dataset is treated as bag of terms or words along with its POS tags. Each term in the bag has been associated a weight which is also considered as the number of times the term appeared in the document or in other words the term frequency in the document. By associating the frequency of each term we can remove redundant terms, which further helps in reducing the size of the information set.

B. Sentiment Analysis

Sentiment analysis or opinion mining is the task of identifying and extracting subjective information of opinions and then categorizing opinions into different classes based on the subjectivity often “positive" , “negative", or “neutral”. Generally, the goal of sentiment analysis is to determine the judgment of a speaker or writer on a topic or the overall subjective polarity of a document [15]. We have used polarity scores of terms to determine the subjectivity. Polarity of terms are determined with the help of an opinion lexicon resource. For this purpose we have used SentiWordNet 3.0 [22].

Opinion lexicons are resources that associate each synset (w) (a set of synonyms) of WordNet with three numerical scores Obj(w) (how objective the word is), Pos(w) (how positive the word is), and Neg(w) (how negative the word is). Their utilization in opinion analysis research comes from the hypothesis or speculation that individual words could be considered as a unit of assessing the opinion or information. They may provide some information about subjectivity of the data. In SentiWordNet each score ranges from 0 to 1, and the sum of these scores always equals to 1 for each WordNet synset. The range of polarityScore(w) always lies between -1 to 1. Score equal to 0 indicates neutral word, score close to -1 indicates highly negative term and score close to +1 indicates highly positive term [10].

C. Feature Engineering

A feature can be defined as a piece of information about the data or text which can be useful in making some predictions [1]. In order to perform classification, we proposed a number of features on the basis of polarity scores of terms available in the sentiments. A sentiment refer to the text record of comment or opinion. These features provide an approximation of positive/negative subjectivity of sentiments. The following is the explanation of features used:-

Given a sentiment s (opinion/comment), let t represent a term in the sentiment s.

- ***Sum of Polarity scores:*** This model returns T , which is the sum of polarity scores of all terms t in the given sentiment s and it is defined as:

$$T = \sum_{t \in s} polarityScore(t) \tag{1}$$

- ***Sum of Positive Polarity scores:*** This model returns +T, which is the sum of polarity scores of only those terms that have $polarityScore(t) > 0$. Here$+t$ represents the set of all positive terms in the given sentiment s , the equation is defined as:

$$+T = \sum_{+t \in s} polarityScore(+t) \tag{2}$$

- ***Sum of Negative Polarity scores:*** This model returns -T, which is the sum of polarity scores of only those terms that have $polarityScore(t) < 0$. Here$-t$ represent the set of all negative terms in the given sentiment s, the equation is defined as:

$$-T = \sum_{-t \in s} polarityScore(-t) \tag{3}$$

- ***Count of Positive Terms:*** This model returns $+CT$ the count of positive terms ($|+t|$) in s that have $polarityScore(t) > 0$ and it is defined as:

$$+CT = \sum_{+t \in s} |+t| \tag{4}$$

- ***Count of Negative Terms:*** This model returns $-CT$ the count of negative terms ($|-t|$) in s that have $polarityScore(t) < 0$ and it is defined as:

$$-CT = \sum_{-t \in s} |-t| \tag{5}$$

- ***Ratio of Terms:*** This model returns R the ratio of all terms in s to positive terms $+CT$ (that have $polarityScore(t) > 0$) with respect to sentiment length and it is defined as:

$$R = \frac{+CT}{|s|} \tag{6}$$

D. Machine Learning

The extracted features from the data set are used to classify the opinions as positive or negative. We have used two the state of the art classification algorithms, such as, naive Bayes classifier (NB) and decision tree (J48) [9]. The motivation behind using two different algorithms for experiments was to evaluate whether the higher accuracy is achieved due to the classification algorithms or engineered features have good descriptive power to accurately classify positive/negative content.

Naive Bayes Classifier (NB): Naive Bayesian classifier is a simple probabilistic classifiers. It is a famous classifier for text categorization. Naive Bayesian classifier uses posterior probability through Bayes formula to classify a sample of document. It is a simple but effective classifier which presumes that the presence or absence of a specific feature is unrelated to the presence or absence of any other feature, given the class variable. Documents are classified by computing their probability of being in any one of the given categories, and assigned to the category having highest probability [24]. With appropriate preprocessing, it's performance is comparable with more advanced classifiers including support vector machines and neural networks. One major advantage of Naive Bayesian classifier is that it only requires a small set of training data to classify unknown samples [2].

Decision Tree (J48): J48 classifier is an extension of decision tree classifier. J48 recursively partitions the prediction space to model the relationship between features and classes. Using a set of training samples a tree is constructed. J48 adopts a top-down approach that searches a learning model in a part of the search space. Traversing the resultant tree gives a set of rules that can be used for classifying unknown samples into the given classes [2].

IV. EVALUATION

A. Experiment Settings

We conducted our experiments with two different available data-sets for this research. The first data-set [20] contains text such as video titles, description and comments given by different users on YouTube, which were collected by various researchers after crawling of the website. Later on, from that data-set we chose around 200 different video titles/comments and carried out an on-line survey. The survey involved 20 participants and helped us in mining the opinions of the people, and labeling or categorizing the text as positive or negative. The text was then parsed through the different document pre-processing steps and then was semantically analyzed to be ready for machine learning. The second data-set [11] contained comments from different social sites. The comments were already classified as negative and positive classes. We selected 3875 comments randomly from the testing corpus to test the features, we engineered for classification.

B. Experimental Results

We used WEKA a popular machine learning tool to train both NB and J48 classifiers [9]. We found that these 2 algorithms performed competitively for classifying the text into two different classes; Positive and Negative.

The data was classified by selecting one feature at a time and as well as all the features together at once. The performance of the both classifiers, as shown in Table I, remained in the range of 54.3% to 71.4% accurate for the dataset acquired from YouTube. The highest accuracy is achieved by the features "Count(+Terms)" and "Ratio Of Terms" using J48 classifier, which is of 71.35%. Whereas, we see that the feature "Sum Of All Terms" provided the minimum accuracy of 61.8% using J48 classifier. Naive Bayes classifier generated slightly lower accuracy results as compared to J48. The highest accuracy of 68.3% is achieved through the feature "Count of positive terms" and the lowest accuracy of 54.3% is achieved through feature "Ratio of Terms" using Naive Bayes classifier. Since the algorithms of the classifiers are heuristics based, so it is obvious to see some differences between the accuracies of the classifiers used in experiment.

To evaluate our experiment results, we used a second dataset, which contained already classified comments. We selected 3875 comments randomly from the testing corpus to test the features, we engineered for classification. The performance of classifiers, as shown in Table II, varied between the range of 71.7% to 73.2% for the dataset used to evaluate the system. The highest accuracy is achieved by the 3 features "Sum of All Terms", Count(+Terms) and "Ratio Of Terms", using J48 classifier, which is of 73.1613%. Whereas, we see that the feature "Count(-Terms)" provided the minimum accuracy of 71.69% using J48 classifier. Here the performance of Naive Bayes classifier remained very closer to J48 classifier in terms of accuracy. The highest accuracy of Naive Bayes classifier was also 73.1613% and the lowest accuracy of 71.69%. These evaluation accuracy results on evaluation dataset were 2%-3% higher than the results obtained through 1st experimental dataset. Thus, providing better evaluation results as compared to the experimentation results. Moreover, when comparing the two classifiers, both classifiers generated results very close

Total # records: 199			Sum (All Terms)	Sum (+Terms)	Sum (-Terms)	Count (+Terms)	Count (-Terms)	Ratio (+/All Terms)	All Features
J48 Classifier	Accuracy	Correctly Classified Records	123 61.809 %	132 66.3317%	132 66.3317%	142 71.3568%	131 65.8291%	142 71.3568%	127 63.8191%
		Incorrectly Classified Records	76 38.191 %	67 33.6683%	67 33.6683%	57 28.6432%	68 34.1709%	57 28.6432%	72 36.1809%
Naïve Bayes Classifier	Accuracy	Correctly Classified Records	123 61.809 %	125 62.8141%	133 66.8342%	136 68.3417%	131 65.8291%	108 54.2714%	135 67.8392%
		Incorrectly Classified Records	76 38.191 %	74 37.1859%	66 33.1658%	63 31.6583%	68 34.1709%	91 45.7286%	64 32.1608%

Table I
CLASSIFICATION ACCURACY RESULTS OF YOUTUBE DATASET

Total # records: 3875			Sum (All Terms)	Sum (+Terms)	Sum (-Terms)	Count (+Terms)	Count (-Terms)	Ratio (+/All Terms)	All Features
J48 Classifier	Accuracy	Correctly Classified Records	2835 73.1613%	2830 73.0322%	2825 72.9032%	2835 73.1613%	2778 71.6903%	2835 73.1613%	2828 72.9806%
		Incorrectly Classified Records	1040 26.8387%	1045 26.9677%	1050 27.0967%	1040 26.8387%	1097 28.3097%	1040 26.8387%	1047 27.0194%
Naïve Bayes Classifier	Accuracy	Correctly Classified Records	2835 73.1613%	2830 73.0322%	2797 72.1806%	2835 73.1613%	2808 72.4645%	2791 72.0258%	2778 71.6903%
		Incorrectly Classified Records	1040 26.8387%	1045 26.9677%	1078 27.8194%	1040 26.8387%	1067 27.5355%	1084 27.9742%	1097 28.3097%

Table II
CLASSIFICATION ACCURACY RESULTS OF EVALUATION DATASET

to each other. Only difference of 0.0% to 1.4% accuracy is seen between the two classifiers. The results have shown that engineered features have good descriptive power to accurately classify positive/negative content.

Since text shared on social platforms contain lots of abbreviations as people try be as short as possible to save time and also there are higher chances of spelling mistakes, thus it has been hypothesized that it is very difficult to come up with very high accuracy results; until and unless these spelling mistakes and abbreviation problems are catered while performing document pre-processing tasks.

V. CONCLUSION & FUTURE CHALLENGES

In this paper we proposed a novel text based content filtering approach for the filtering or blocking of web pages on the basis of sentiment analysis of the text present on web pages. Our methodology contained techniques from information filtering, opinion mining, sentiment analysis, and machine learning for efficient filtering of content available on the websites. By choosing the best techniques from different areas, we overcame the drawbacks faced through typical URL filtering and keyword matching techniques. The results show that the filtering techniques based on sentiment analysis and opinion mining combined with machine learning classification, perform well enough to be worth studying further as a research topic.

The major challenge of this approach is to come up with techniques which can help reduce noise from text, as the text present on social platforms are mostly prone to spelling mistakes and contain many abbreviations and emotion symbols. Therefore, improvements in the accuracy of classification can be achieved by combining the techniques such as spell correction, use of abbreviation matching dictionary and emotion symbols detection to reduce noise from text available on social websites. Furthermore, updating the SentiWordNet dictionary to obtain the sentiment scores of maximum words in the text can also play a major role in achieving highly accurate results.

REFERENCES

[1] *Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus*. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM., 2012.

[2] Aggarwal, C. C, and C. Zhai, *Mining text data*, USA, 2012, ch. A survey of text classification algorithms, pp. 163–222.

[3] A. Bhutani, D. K. Misra, and S. Toshniwal, *Detecting Socially Offensive Comments*, Indian Institute of Technology, Khanpur Std., November 2012.

[4] Y. Chen, S. Zhu, Y. Zhou, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety." 2012 International Conference on Social Computing (SocialCom). IEEE, 2012, 2012, pp. 71–80.

[5] J. Cheng., *Report: 80 percent of blogs contain offensive content*, Std., 2007. [Online]. Available: http://arstechnica.com/security/news/2007/04/report-80-percent-ofblogs-contain-offensive-content.ars

[6] N. Churcharoenkrung, Y. S. Kim, and B. H. Kang, "Dynamic web content filtering based on users knowledge," 2005.

[7] R. Du, R. Safavi, and W. Susilo, *Web filtering using text classification*, Faculty of Informatics-Papers (2003): 166. Std., 2003.

[8] B. Guc, "Information filtering on micro-blogging services," Master's thesis, Swiss Federal Institute of Technology Zurich, August 2010.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H, *Weka: Data mining Software in Java*, Univeristy of Waikato Std., 2009. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[10] A. Hamouda and M. Rohaim, "Reviews classification using sentiwordnet lexicon," *The Online Journal on Computer Science and Information Technology (OJCSIT)*, vol. 2, pp. 120–123, 2011.

[11] K. Inc, *Detecting insults in social commentary*, Std., 2012. [Online]. Available: http://www.kaggle.com/c/detecting-insults-in-social-commentary/data

[12] P. Karampiperis, A. Koukourikos, and G. Stoitsi, "Collaborative filtering recommendation of educational content in social environments utilizing sentiment analysis techniques," *Recommender Systems for Technology Enhanced Learning: Research Trends & Applications*, vol. RecSysTEL Edited Volume, Spinger, 2013.

[13] H. Khandelwal, "Polarity detection in movie reviews."

[14] P. Lee, S. Hui, and A. Fong, "A structural and content based analysis for web filtering," *Internet Research: Electronic Networking Applications & Policy*, vol. 13, pp. 27–37, 2003.

[15] B. Liu, "Sentiment analysis and opinion mining," Morgan & Claypool Publishers, May 2012.

[16] A. K. McCallum, *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*, Std., 1996. [Online]. Available: http://www.cs.cmu.edu/ mccallum/bow

[17] J. Palme, *Information Filtering*, Std., 2001. [Online]. Available: http://people.dsv.su.se/ jpalme/select/information-filtering.pdf

[18] R. S. Ranjani and T. Sheela, "Impact of video recommendation system and filtering technique on dissemination of polluted content," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, January 2013.

[19] L. Ruibo and X. Hao, "Research and implementation of bho-based content filtering system." 2013 International Conference on Information Science and Technology Application (ICISTA-13), Atlantas Press, 2013.

[20] S. F. I. (SFI), *Insight Project Resources*, SFI Insight Centre for Data Analytics Std., 2012. [Online]. Available: http://mlg.ucd.ie/index.html

[21] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, *How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings*, The International World Wide Web Conference Committee (IW3C2) Std., April 2010.

[22] B. Stefano, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In LREC, 2010, vol. 10, pp. 2200–2204.

[23] K. Toutanova, *Stanford Log-linear Part-Of-Speech Tagger*, The Stanford Natural Language Processing Group, Stanford University Std., 2000. [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml

[24] Wikipedia. (2014) Bag of words models. [Online]. Available: http://en.wikipedia.org/wiki/Bagofwordsmodel

[25] ——, *Tokenization*, Wikimedia Foundation, Inc Std., June 2014. [Online]. Available: http://en.wikipedia.org/wiki/Tokenization

[26] Z. Zhou, T. Song, and Y. Jia, "A high-performance url lookup engine for url filtering systems," IEEE. 2010 IEEE International Conference on Communications (ICC), May 2010, pp. 1–5.