# Stop the KillFies! Using Deep Learning Models to Identify Dangerous Selfies

Vedant Nanda*, Hemank Lamba§, Divyansh Agarwal*, Megha Arora§, Niharika Sachdeva*, Ponnurangam Kumaraguru*

*IIIT-Delhi, § Carnegie Mellon University
{vedant15114,divyansha,niharikas,pk}@iiitd.ac.in,{hlamba}@cs.cmu.edu,{marora}@andrew.cmu.edu

## ABSTRACT

Selfies have become a prominent medium for self-portrayal on social media. Unfortunately, certain social media users go to extreme lengths to click selfies, which puts their lives at risk. Two hundred and sixteen individuals have died since March 2014 until January 2018 while trying to click selfies. It is imperative to be able to identify dangerous selfies posted on social media platforms to be able to build an intervention for users going to extreme lengths for clicking such selfies. In this work, we propose a convolutional neural network based classifier to identify dangerous selfies posted on social media using only the image (no metadata). We show that our proposed approach gives an accuracy of 98% and performs better than previous methods.

**ACM Reference Format:**
Vedant Nanda*, Hemank Lamba§, Divyansh Agarwal*, Megha Arora§, Niharika Sachdeva*, Ponnurangam Kumaraguru*. 2018. Stop the KillFies! Using Deep Learning Models to Identify Dangerous Selfies. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3191575

## 1 INTRODUCTION

A selfie is defined as *a photograph that one has taken of oneself, typically taken with a smartphone or a webcam and shared via social media* [25]. The popularity of selfie culture can be estimated from the fact that in 2015, 24 billion selfies were uploaded to Google Photos [1]. Pew research reported that around 55% of millennials have posted a selfie on a social media platform [3]. Selfie nowadays has become a ubiquitous tool for self-presentation on social media.

Previous research has extensively focussed on the psychological and social variables of the people who post selfies. These works show that people posting a lot of selfies have personality traits such as narcissism, lack of self-esteem, self-embellishment and social alienation [8, 12]. Self-embellishment has been reported as one of the primary reasons for clicking a selfie; most selfies are clicked to be posted on a social platform [1]. In extreme cases, users may often engage in dangerous activities and situations to click selfies which might make them popular on social media [8, 13]. Users often engage in such situations to portray themselves as adventurous and enhance their appearance to others while risking their own physical well-being [9, 18]. Continuing the statistic in [16], we found that as many as 216 individuals have died while attempting to take selfies.

We define a dangerous selfie as *a selfie which potentially might cause harm to an individual or a group that may occur while the individual(s) attempts to take a selfie.* To be able to detect the users who post such dangerous selfies, and to make an intervention, it is essential to find and identify dangerous selfies. By identifying such selfies being posted on the social media platform by a user, combined with the frequency at which the user is posting them, the social networking platform can decide if a particular user is overindulging in risk-taking behavior, which could potentially be harmful to their health. In this work, we propose a deep-learning based framework to identify dangerous selfies posted on Twitter. We use existing deep neural networks such as VGG16 and VGG19 [20], Inception v3 [23], ResNet50 [11] etc. and adapt it to perform well on the task of detecting dangerous selfies. We discover that our model outperforms the previously proposed models by a factor of 1.34 in terms of accuracy on the test set. We believe that this work will help researchers understand a user's propensity to post such selfies on online social media in a much better way, thus resulting in effective intervention technologies.

## 2 RELATED WORK

Numerous research works have investigated the effect of selfie culture on the mental well-being of selfie-er. Researchers discovered that the people who post more selfies have shallow relationships with people [12] or decreased intimacy [2], ultimately leading to feelings of loneliness and worry. These users were also found to have the dark triad personality (narcissism, psychopathy, and machiavellianism) [8]. Previous research has also tried to view the number of likes, comments, and shares an individual gets for their selfies as the social currency for the youth, and this desire of gaining more of such currency prompts youth to extreme lengths [15].

**Selfies and physical harm**: Subrahmanyam et al. [21] discuss how selfie can cause physical harm to the selfie-ers in different situations. Lamba et al. was the first work in the area of dangerous selfies to characterize the number of selfie deaths in the past years, and analyze their victims and causes [16]. They also proposed a multi-modal classifier which takes into account posts' text, image, and location to identify if a particular user is in a dangerous situation or not. In this work, we show how our method outperforms the previously proposed approaches.

[1] https://googleblog.blogspot.in/2016/05/google-photos-one-year-200-million.html


*WWW '18 Companion, April 23–27, 2018, Lyon, France*

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191575

**Deep-Learning and Image Recognition**: In the recent years, a lot of work has been done in the field of large-scale visual recognition and image classification. Many methods are available, including Alexnet [14], which was the first model to popularize the use of convolutional neural networks (CNNs) for object recognition. Following AlexNet, many different architectures were proposed and improvement was noted with GoogLeNet/Inception [22], VGG 16 and VGG 19 [20], and Inception v3 [23]. All of these architectures have obtained high accuracies of classifying images in the Imagenet dataset. However, training and testing them on social media images has been a challenge since getting annotations isn't easy.

The past work on classifying dangerous selfies is heavily dependent on using image captions, text and location features of a post. Leetaru et al. [6] find that only 1.6% of tweets have the exact location, which makes it hard to infer using the previously proposed model if a given selfie is dangerous or not. Moreover, getting image captions can also be challenging in some cases (say on a smartphone), thus rendering the previously proposed approach non-tractable. Our method leverages the high learning capabilities of these very deep neural networks to build a classifier which detects if a given image is a dangerous selfie or not. We also try SVMs and fine-tune the model to perform a large-scale analysis of how the models perform for our task.

## 3 DATASET

For the data-collection process, we chose Twitter as it is a popular social media observing selfie culture. We collected tweets related to selfies by searching words like *selfie* or its immediate variants (#selfie, #dangerousselfie, #extremeselfie, #letmetakeaselfie, #selfieoftheday, and #drivingselfie). The data collection was done between August 1, 2016, and September 27, 2016. Through this method, we obtained 138K unique tweets posted by 78K individual users. The dataset was filtered for only images and geo-location. Following this, we were left with 9,444 geocoded tweets. To validate which of the images contained in tweets were selfies, we trained a classifier (explained below).

**Pre-processing:** We used the same preprocessing methodology as proposed in [16]. We use a classifier to distinguish selfie images from non-selfie images based on the CNN model architecture InceptionV3 [23]. After curating, we manually annotate a dataset of 2.1K images into 1.3K selfies and 800 non-selfies. We used a transfer learning framework (DeCAF [5]) to retrain the Inception model for our dataset. We found that this model gave 88.48% accuracy with 10-fold cross-validation using which the labels (selfie or not a selfie) were obtained for all the 9,444 geocoded images. This process yielded a candidate set of 6,842 tweets which were potential tweets containing selfies, rest being flagged by the model as non-selfie tweets.

**Manual Annotation:** The final step for identifying dangerous selfies involved human annotations on the obtained selfie candidate set of 6,842 tweets. For the purpose of annotation, we developed a web interface and provided each annotator with an authenticating login and password. We recruited annotators via posting a request for participation on the mailing list of different universities. The annotation session started with a 15 minute introduction about the annotation procedure. All annotators used the "dangerous selfie definition" provided by the authors in Section 1. Following the introduction, each annotator marked whether they would consider the shown image as a selfie and if so, whether it is a dangerous selfie or not. We also asked annotators to note the possible reason for it being dangerous such as "selfie was taken on a mountain". Each selfie was annotated by 3 distinct annotators. The inter-annotator agreement rate, using the Fleiss Kappa metric [7] was 0.58, thus indicating moderate agreement between the annotators [17]. We used majority voting to decide the final label for a given selfie, and ties were resolved randomly. We found that from the selfie candidate set of 6,842 tweets, our annotators agreed that 6,460 tweets contained selfies. Among these, 623 were marked as dangerous selfie containing tweets and remaining 5, 837 as non-dangerous. We conduct all our future analysis on this set of 6, 460 annotated tweets. It should be noted that this dataset was curated by the authors in a previous work [16] and this work uses the same dataset. [2]

## 4 PROPOSED CLASSIFIER

Previous research showed that multimodal features can be useful for identifying posts containing dangerous selfies on Twitter [16]. Authors showed that the image features (dense captions created by text from the images) gave the best accuracy among all the modes of features. It was also noted that combination of all the three features performed the best, and gave 73% accuracy. In this work, we propose a CNN-based architecture that works only on image-based features. In our work, we leverage the existing deep-learning models that have performed well in identifying images on large-scale benchmark datasets such as Imagenet [19]. These state of the art models are pre-trained on Imagenet dataset used for ILSVRC (Imagenet Large Scale Visual Recognition Challenge), containing 1.2 million images labeled with 1,000 class labels. Applying the same architecture to our dataset, and re-training the network is a challenge, as for successful training of large architecture, a huge number of samples is required, which in our case isn't available. Therefore, we use pre-trained architectures and apply transfer learning for solving our task. We use the weights of models that do well on the Imagenet dataset to fine tune them for our problem statement. The intuition behind doing this is that the image features a model trained on Imagenet is using should be similar to the features we require. We model the problem as a two-class classification problem with the positive class consisting of dangerous selfies (623 samples) and negative class of non-dangerous selfies (5,837 samples).

**Handling Skewness**: The number of positive samples (623 dangerous selfies) in our dataset is much less than the number of negative samples (5,837 non-dangerous selfies). This can be viewed as a rare class classification problem. To have more representative class balance in our dataset, we use data augmentation operations - shift (shifting the image pixels linearly in a range of 20% of width and height of image), flip (flipping the image pixels horizontally and vertically), rotate (rotating the image by a certain degree, randomly chosen in a range of 0-180), and shear (with a random zoom range and shear intensity of 0.2). Following which, we further downsample our dataset to give us a balanced dataset of 3, 115 images in each class. Downsampling is a well-known method to handle class imbalance challenge in classification [10].

[2] http://precog.iiitd.edu.in/requester.php?dataset=killfie2018

**Table 1: Results for fine-tuned models. ResNet50 with 128 nodes in the densely connected layer outperforms other models.**

| Model Name (optimal nodes in the dense layer) | Train set accuracy | Test set accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| VGG 16 (512) | 0.973 | 0.979 | 0.971 | 0.989 | 0.980 |
| VGG 19 (256) | 0.969 | 0.976 | 0.963 | 0.992 | 0.977 |
| InceptionV3 (1024) | 0.965 | 0.962 | 0.945 | 0.985 | 0.965 |
| Xception (2048) | 0.970 | 0.977 | 0.975 | 0.980 | 0.978 |
| **ResNet50 (128)** | **0.979** | **0.981** | **0.982** | **0.982** | **0.982** |
| InceptionResNetV2 (512) | 0.967 | 0.974 | 0.964 | 0.986 | 0.975 |

**Data pre-processing**: Once we obtained 3,115 images in both classes (total 6,230 images), we scaled each image to a size of 224 by 224 pixels and all these images were shuffled to ensure there's no bias in training data. A random 80:20 train test split was then done to get 4,984 and 1,246 images in the train and test respectively. Finally, all images were normalized using the mean and standard deviation of the dataset it was pre-trained on (Imagenet).

**Feature Extraction**: Since our dataset is relatively different from the Imagenet dataset (which is a more generic dataset, and does not limit itself to just selfies), we also explore using the pre-trained models as feature extractors and then fitting a non-neural network based classifier (like SVM) on those features. For feature extraction, we took a pre-trained model and removed the softmax layer. The vector obtained on a forward propagation of an image (the layer just before softmax output) was treated as the feature vector for that image. So if, for example, we use VGG 16 or VGG 19 to extract image features, we get 4,096 features corresponding to each image. ReLu (Rectified Linear unit) activation was applied to these features which were then used to train SVMs with linear and RBF kernels.

**Training SVMs**: We use the primal formulation of SVM and train both soft and hard margin classifiers by tuning the hyperparameter C where a larger value of C corresponds to more penalty for misclassification and a smaller C in lesser penalty thus leading to hard and soft margin classifiers respectively. We find the best value of C by doing grid search for $C \in [0.01,1.28]$ where step size of the interval was increased exponentially with each iteration. Accuracy was used as the metric to evaluate the best value of C using 3 fold cross-validation on the training set. We also explore the application of Principal Component Analysis (PCA) [26] to the extracted features using which we reduce the dimension of each feature vector to 100. Further both linear and Radial Basis Function(RBF) were used as kernel functions. For each of the architectures used for feature extraction and corresponding to each kernel function, we get a separate SVM model.

**Architecture Customization**: The Imagenet dataset has $1,000$ classes, and all architectures use the softmax layer as the final layer to make predictions. We modify the architecture by removing the softmax layer from each of the pre-trained models. Further, we apply the global average pooling operation on the output layer. We append this architecture by adding two dense layers - the first one is with ReLu activation function, followed by a layer consisting of 2 nodes with softmax activation, and this becomes our output layer. The number of nodes in the dense ReLu activation layer is treated as a hyperparameter for each model and the best number was decided by applying a grid search using 3-fold Cross-validation.

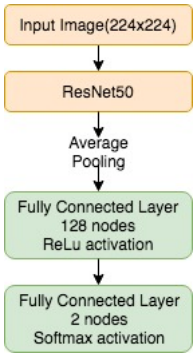


**Figure 1: Proposed Model. We used 3-fold Cross-validation and found that 128 nodes in the dense layer gives best Cross-validation accuracy.**

For training, all the pre-trained layers were frozen, and weights of densely connected layers - which were initialized randomly - were trained for 300 epochs with a batch gradient descent, keeping batch size to be 50. After training the densely connected layers, last two layers of the pre-trained model were unfrozen and fine-tuned along with the densely connected layers for another 300 epochs. We experimented with the following architectures - VGG-16, VGG-19, InceptionV3, Xception, ResNet50, and InceptionResNetV2 [4, 11, 20, 23, 24]. The results for these models, along with the best hyperparameter - which in this case is the number of nodes in the densely connected layer - are presented in Table 1 and are discussed in Section 5. The proposed architecture is shown in Figure 1.

## 5 RESULTS

To ensure that these results are not a false indication of the performance of the models, we make sure at the time of train-test split that data is properly shuffled. This results in a fairly balanced test set containing 649 and 597 samples in the positive and negative class respectively. Other than test set accuracy, we also make sure we look at other factors such as training accuracy, precision, recall and F1 score to make sure the model hasn't overfitted and the insights obtained are correct.

### 5.1 Feature Transformation

From Table 2, we see that features extracted from ResNet50 work better than other models giving a test set accuracy of 95% with PCA and RBF kernel. High precision, recall, and F1 scores further validate the model's performance. Another interesting thing to note is that

**Table 2: Results for SVM with and without PCA. Features extracted using ResNet50 perform best for both linear and RBF kernels, both with and without PCA. Overall, PCA along with RBF kernel performs best.**

| | **Linear Kernel (with PCA)** | | | | | **RBF Kernel (with PCA)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Feature Extractor** | **Train set acc** | **Test set acc** | **Precision** | **Recall** | **F1 score** | **Train set acc** | **Test set acc** | **Precision** | **Recall** | **F1 score** |
| VGG 16 | 0.876 | 0.860 | 0.872 | 0.858 | 0.865 | 0.991 | 0.601 | 1.000 | 0.234 | 0.380 |
| VGG 19 | 0.888 | 0.870 | 0.900 | 0.844 | 0.871 | 0.980 | 0.600 | 1.000 | 0.233 | 0.378 |
| Inception V3 | 0.903 | 0.888 | 0.915 | 0.866 | 0.890 | 0.997 | 0.921 | 0.951 | 0.894 | 0.921 |
| Xception | 0.915 | 0.908 | 0.929 | 0.891 | 0.910 | 0.966 | 0.932 | 0.958 | 0.909 | 0.933 |
| **ResNet50** | **0.938** | **0.937** | **0.961** | **0.917** | **0.938** | **0.999** | **0.952** | **0.990** | **0.917** | **0.952** |
| InceptionResNetV2 | 0.916 | 0.903 | 0.918 | 0.894 | 0.906 | 0.994 | 0.941 | 0.944 | 0.941 | 0.943 |
| | **Linear Kernel (without PCA)** | | | | | **RBF Kernel (without PCA)** | | | | |
| **Feature Extractor** | **Train set acc** | **Test set acc** | **Precision** | **Recall** | **F1 score** | **Train set acc** | **Test set acc** | **Precision** | **Recall** | **F1 score** |
| VGG 16 | 0.999 | 0.914 | 0.903 | 0.935 | 0.919 | 0.987 | 0.927 | 0.927 | 0.934 | 0.930 |
| VGG 19 | 0.999 | 0.898 | 0.894 | 0.912 | 0.903 | 0.981 | 0.919 | 0.924 | 0.920 | 0.922 |
| Inception V3 | 0.967 | 0.911 | 0.931 | 0.895 | 0.913 | 0.950 | 0.914 | 0.943 | 0.889 | 0.915 |
| Xception | 0.960 | 0.934 | 0.949 | 0.923 | 0.936 | 0.926 | 0.911 | 0.941 | 0.884 | 0.912 |
| **ResNet50** | **0.982** | **0.941** | **0.960** | **0.924** | **0.942** | **0.961** | **0.941** | **0.974** | **0.911** | **0.941** |
| InceptionResNetV2 | 0.958 | 0.922 | 0.939 | 0.909 | 0.924 | 0.949 | 0.924 | 0.941 | 0.911 | 0.926 |

ResNet50 works better regardless of the kernel or feature selection using PCA. This indicates that ResNet50 is a better feature extractor than other architectures when it comes to the task of detecting dangerous selfies. We posit that ResNet's ability to classify objects in the Imagenet dataset with higher accuracy (top-5 validation error of 6.71% compared to VGGnet's 8%) than other models is a contributing factor as Imagenet contains images having objects like vehicles, animals/insects, scenes of a cliff, water body etc. presence of which can possibly result in a dangerous selfie.

## 5.2 CNN-based Results

We see from Table 1 that even for a fine-tuned models, ResNet50 gives the best performance. It achieves a test set accuracy of 98% getting high (0.98) precision and recall values. The training set accuracy (97.9%) is slightly lower than test set accuracy thus showing that the model hasn't overfitted and generalizes well. Overall fine-tuned models perform best in terms of all the evaluation metrics (precision, recall, accuracy). We further show that deep-learning approaches perform really well obtaining high accuracy, precision, and recall over the test set. It also performs better than the previous classifier proposed in the literature (Fig 2).

## 6 CONCLUSION

In this work, we conduct a large scale analysis of machine learning models to classify an image as a dangerous or a non-dangerous selfie. Our work shows that using only image features can identify dangerous selfies more accurately than using the multimodal features, as proposed in existing literature. We achieve an accuracy of 98% with high precision and recall values, as compared to the 73% accuracy achieved previously. This is a major improvement not only in terms of accuracy but also in terms of usability. Using only image features does not require image captions, posts' location or text to classify an image – some of which might not be available in certain cases. It should also be noted that using only the image features for the classification task allows real-time detetection and can be used to build effective intervention technologies. Thus, we provide a more usable, robust and an accurate solution.

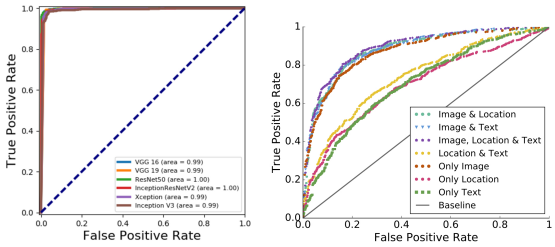


**Figure 2: ROC curve for fine-tuned models on the left compared to previous models. Fine-tuned models show high area under the curve thus showing robust classification capabilities.**

## 7 DISCUSSION

While this work explores and proposes a classification framework for an important problem of identifying dangerous selfies, there is definitely an immediate need to apply this classifier in real world scenarios. Further tools need to be developed over the classifier, which can exactly identify individuals at risk and also build efficient intervention tools which can potentially prevent deaths and injuries. We have already built a crowdsourcing tool called "Saftie", which is available on app store (goo.gl/2sIdYT) and as a Facebook chatbot (fb.me/saftiebot). This tool is currently live, and has collated about 1500+ dangerous locations, where a person should not click selfies. Another tool, currently in development, is to integrate the classifier into the camera app, which can nudge a user in an effective way before they try to click a dangerous selfie. We hope this work can result in more such technologies, that can prevent further harm.

## REFERENCES

[1] Shivani Arora. 2016. Social Networking-A Double-Edged Sword. *International Conference On Recent Trends In Engineering Science And Management* (2016).
[2] L. Blaine. 2013. How Selfies Are Ruining Your Relationships. (2013).
[3] Pew Research Center. 2014. More than half of Millennials have shared a 'Selfie'. (2014).
[4] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR* abs/1610.02357 (2016).
[5] J. Donahue et al. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *ICML*.
[6] Kalev Leetaru et al. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18, 5 (2013). https://doi.org/10.5210/fm.v18i5.4366
[7] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* (1971).
[8] Jesse Fox and Margaret C Rooney. 2015. The Dark Triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences* (2015).
[9] Jennifer Guay. 2014. Most dangerous selfies. (2014).
[10] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Trans. on Knowl. and Data Eng.* 21, 9 (Sept. 2009), 22.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015).
[12] David Houghton et al. 2013. Tagger's delight? Disclosure and liking in Facebook: the effects of sharing photographs amongst multiple known social circles. (2013).
[13] Peter K. Jonason and Laura Krause. 2013. The emotional deficits associated with the Dark Triad traits: Cognitive empathy, affective empathy, and alexithymia. *Personality and Individual Differences* 55, 5 (2013), 532 – 537.
[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. 1097–1105.
[15] AK Lakshmi. 2015. The Selfie Culture: Narcissism or Counter Hegemony? *Journal of Communication and media Studies* (2015). Issue 1.
[16] Hemank Lamba et al. 2017. From Camera to Deathbed: Understanding Dangerous Selfies on Social Media. (2017).
[17] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977).
[18] Mark R Leary et al. 1994. Self-presentation can be hazardous to your health: impression management and health risk. *Health Psychology* (1994).
[19] Olga Russakovsky et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision* 115, 3 (2015).
[20] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).
[21] BV Subrahmanyam and et al. 2016. Selfie Related Deaths Perils of Newer Technologies. *Narayana Medical Journal* (2016).
[22] Christian Szegedy et al. 2014. Going Deeper with Convolutions. *CoRR* abs/1409.4842 (2014).
[23] Christian Szegedy et al. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015).
[24] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR* abs/1602.07261 (2016).
[25] I Taslim and Md Z Rezwan. 2013. Selfie Re-de-fined: Self-(more/less). *Wizcraft Journal of Language and Literature* (2013).
[26] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1 (1987), 37 – 52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.