Evaluation of Mental Health Classification with Social Media

CHARLES INWALD, Lehigh University

This study compares several techniques in studying mental health on the social network Reddit. Combinations of different feature selections and classification models are evaluated in terms of accuracy on classifying text as pertaining to Reddit's communities for depression, anxiety and suicidal ideation. Special consideration is given to whether techniques are likely to understate the severity of cases of suicidal ideation as depression.

Additional Key Words and Phrases: topic modeling, social networks, natural language processing

ACM Reference Format:
Charles Inwald. 2019. Evaluation of Mental Health Classification with Social Media. 1, 1 (January 2019), 9 pages.

1 INTRODUCTION

The World Health Organization estimates that in the year 2020, the world will average one suicide every 20 seconds and one suicide attempt every one to two seconds [1]. Many are concerned by the rise in suicide rates, with the United States suicide rate rising over 25% from 1999 to 2016 [2]. While correlation does not necessarily imply causation, it has been found that the rise in popularity of social networks such as Facebook coincide with substantial increases in the national suicide rate, as shown in Figure 1 [2]. Possible explanations for this include the increased connectivity that social media brings, which consequently leads to fear of missing out when one sees their friends without them, cyber-bullying, and unrealistic expectations from comparing one's internal reality to the external facades of others on social media [2].

On the other hand, certain areas social media, and other online resources, serve as a refuge for those struggling with poor mental health. Notable examples of this phenomenon include Reddit's /r/SuicideWatch, /r/depression, and /r/anxiety. These online communities aim to provide peer support, with many of the users who reply to cries for help having similar struggles themselves, according to a moderator thread on /r/depression [3]. The emergence of these online communities containing large corpora of text pertaining to the mental health of their authors, creates a new avenue for risk assessment and classification.

This paper investigates the feasibility of several natural language processing techniques to analyze the troves of public facing corpora of Reddit's mental health communities, in the hopes of contributing towards the goal of a web tool that can accurately flag at risk individuals, allowing for quick identification and response. Specifically, the study investigates the accuracy of classifiers on categorizing Reddit posts into belonging to Reddit's /r/SuicideWatch, /r/depression, and /r/anxiety, using public facing data only. The study also compares the efficacy of the approaches of using topic modelling and word embeddings as feature vectors, and several models for classification.

It is important to note that the techniques presented in this paper, and by extension the hypothetical usage of a tool for identifying at risk individuals, should be used strictly in performing flagging tasks that would not have been performed otherwise. Thus the techniques and tools discussed are not to be used in place of standardized procedures for evaluating at risk individuals, such as the guidelines provided by the National Institute of Mental Health [4].

1.1 Related Work

Significant strides have been made in the area of studying themes in how suicidal ideation manifests in language on social media. There is less work, however, in the niche of feature extraction and

Author's address: Charles Inwald, cci219@lehigh.edu, Lehigh University.

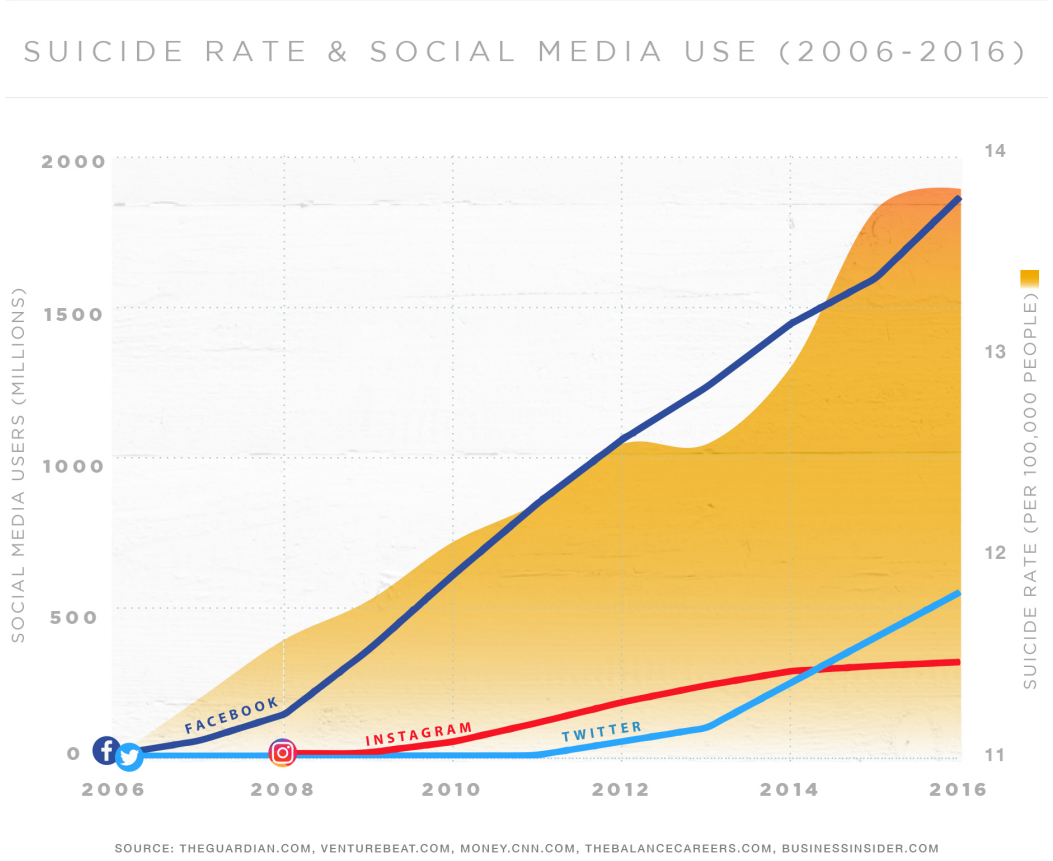

Fig. 1. Suicide Rate and Social Media Use from 2006 to 2016. The rise in popularity of Facebook is correlated with a rise in the national suicide rate. [2]

binary classifiers. De Choudhury et al's paper "Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media", for example, focuses primarily on "language and interactional measures" [5].

Sawhney et. al in "A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets" propose a set of features for use in several binary classifiers [6]. This paper aims to explore a similar objective but with comparison of different feature sets, as well as the added distinction between depression and anxiety.

Shing et. al's paper "Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings" utilized a diverse range of features for their classifier, including topic modeling [7]. This study originally hoped to utilize the dataset from this study, as the clinician and crowdsourced annotations of /r/SuicideWatch posts would provide utility, however it was not acquired within a sufficient timeframe prior to publication. The impact of the individual features on classification accuracy and F2 scores were not discussed, contributions that this study aims to provide.

2 DATASETS

2.1 Data Collection

The input data for this experiment was sourced from Reddit's /r/depression, /r/anxiety, and /r/suicidewatch communities. However, the companion code for this paper can be easily repurposed for alternative use cases by simply replacing these respective variables. There is also the potential adaptation for other social network platforms such as Twitter. However, it should be noted that in both the cases of intra-domain (i.e using alternative subreddits) and cross-domain (i.e using Twitter) adaptation of this work, there may exist instances of the cross-domain subproblem and therefore a possible reduction in performance [8]. In machine learning, we frequently make the assumption of similar distribution of data between the training and target datasets [8]. The cross-domain subproblem exists when we assume this similarity to be the case between domains [8]. The cross domain subproblem is discussed further in Section 8.

It is likely that the magnitude of the classification accuracy reduction would be substantial in extending the model for use on Twitter or other corpora, however exploration of this claim is beyond the scope of this paper. One notable distinction between Reddit and Twitter is the post character length limit, which are 40,000 and 280 characters respectively [10].

The Reddit posts were scraped from Reddit using PRAW for Python [11]. The following sample sizes were used in experimentation. These sample sizes have a standard deviation of 8.05, suggesting that the dataset was sufficiently balanced.

	/r/depression	/r/anxiety	/r/SuicideWatch
Number of Posts	978	962	980

3 PREPROCESSING

The study begins with a preprocessing phase, with the primary purpose of transforming the vast troves of Reddit data into a format adequate for data mining. Preprocessing can be evaluated in terms of memory usage, time complexity, and retrieval performance [14]. This study prioritizes retrieval performance, as scalability is not yet of immediate concern.

3.1 Tokenization

The first stage of preprocessing is tokenization, in which the documents, or in this case posts, are disaggregated into "meaningful units" [12]. While the granularity of this step may vary by use case, for this study, where the inputs are posts less than 40,000 characters, tokens are words, as opposed to sentences or whole documents. In this study, this task also included converting all text to lowercase, and was performed using Gensim's Simple Preprocess utility [13].

3.2 Stopword Removal

Stopwords are language specific functional words, such as prepositions and conjunctions, which tend to carry no significant information for our purposes [14]. This study used Natural Language Toolkit's stopword list of 127 stopwords to remove such instances[15].

3.3 Bigram Model

Language models assign probabilities to sequences of words. N-gram models predict the next word in a sequence, with "bigrams" being instances of just two words and "unigrams" with only one word [16]. It is common for phrases to have meaning beyond the combination of the words they are composed of [17]. The bigram model used pays attention to these nuances by considering whether words appear frequently together only in certain contexts [17].

The following scoring function is used to score the probability of any two words i and j occurring together [17]:

$$score(w_i, w_j) = \frac{count(w_i, w_j) - \delta}{count(w_i) \times count(w_j)}$$

The δ refers to a discounting coefficient, as a frequency threshold. This prevents phrases of very infrequent words being formed [17]. This step of preprocessing is only performed during the topic modeling based approach of this study, with a discounting coefficient of 15.

4 TOPIC MODELING

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning framework which uncovers "hidden topics" in high volume document corpora [18]. Documents become "bag of words", implemented as a vector of word counts [18]. Documents are then represented as probability distributions over some number of topics, with each topic being represented as a probability distribution over a number of words [18]. A Lehigh University study described the process of LDA as follows [18]:

(1) For each document, pick a topic from its distribution over topics.
(2) Sample a word from the distribution over the words associated with the chosen topic.
(3) The process is repeated for all the words in the document.

It is important to note that LDA is unsupervised in that there is no known number of topics to use, and thus the number of topics is a hyperparameter. Fortunately, the optimal number of topics can be refined by evaluation of coherence scores [19]. Prior to the advent of coherence scores, it was common for topics to be judged based on size [19]. Although often times the two metrics are correlated, coherence scores are purported to correspond "well with human coherence judgments and makes it possible to identify specific semantic problems in topic models without human evaluations or external reference corpora" [19]. For this study, each trial calculated at runtime the coherence scores with topic counts from 1 to 50, and the topic number with the highest coherence score was used.

The outputs of this process for this study are the topic distributions for each post, which then become the features for the classifiers.

5 WORD EMBEDDINGS

An alternative approach to topic modeling used in this study was using word embeddings as features for the classifier. Word embeddings utilize neural network language modeling to transform a given corpus into a set of feature vectors [20]. While multiple methods exist, this study makes use of the word2vec software, with the skip-gram with negative sampling model.

6 CLASSIFICATION

This study deals with strictly *binary classification*, that is either distinguishing between depression and suicidality or depression and anxiety. This portion of this study is *supervised* in that posts used are from subreddits of labeled origin, known as the *training data set* in which the classifiers are trained. Logistic Regression and Stochastic Gradient Descent (SGD) linear classifiers are compared in this study.

6.1 Model Selection

Both classifiers have a variety of hyperparameters requiring deliberate fine tuning. In order to optimize classifier performance, this study used SciKit Learn's Cross Validated Grid Search, which

performs an "exhaustive search over specified parameter values for an estimator" [21]. This technique utilizes *k-Fold Cross Validation*, which serves to determine an upper bound to the error of generalization error [22]. This is achieved by splitting the dataset into *k* folds, and training the classifier with first *k - 1* folds, and using the last one as a test set [22]. Benefits include making use of the full dataset iteratively, creating the largest possible test set [23].

In amalgamation with k-Fold Cross Validation, grid search repeatedly performs the cross validation, each time with a different combination of hyper parameters, making note of the generalization error [22].

The following parameters were trialled in the Cross Validated Grid Search: [21]:

- **Loss Function** (applies to SGD only) : Hinge, Logistic Regression, Modified Huber, Squared Hinge, Perceptron, Squared Loss, Huber, Epsilon Insensitive, and Squared Epsilon Insensitive
- **Penalty**: L1, L2, Elastinet, None
- **Max Iterations**: 10, 20, 50, 80, 1000, 10000, 100000
- **Inverse of regularization strength (C)**: Logarithmically spaced from 1 to 10,000
- **Solver** (applies to Logistic Regression only): Newton's Method, Stochastic Average Gradient (SAG) , SAG with L1 (SAGA), Limited Memory Broyden Fletcher Goldfarb Shanno (LMBFGS)

6.2 Stacking

Ensemble Learning is the practice of combining multiple machine learning methods for improved performance [24]. *Stacking* is a particular method of ensemble learning where the output of several base predictors serves as the input to the final meta predictor [24]. In large samples, stacking has been shown to perform at least as well as the best single predictor used [24].

7 RESULTS

Table 1. Optimized f1 Scores (Weighted Averages) for Topic Modelling Approach

k = 5 folds	Depression vs. SuicideWatch	Depression vs. Anxiety
Logistic Regression	0.75	0.77
SGD	0.58	0.74
Stacked Generalization*	0.81	0.70

Table 2. Optimized f1 Scores (Weighted Averages) for Word Embedding Approach

k = 5 folds	Depression vs. SuicideWatch	Depression vs. Anxiety
Logistic Regression	0.34	0.56
SGD	0.40	0.52
Stacked Generalization*	0.42	0.47

*Base Predictors of Random Forest and K-Nearest Neighbor, Meta-Predictor of Logistic Regression

A popular, relatively holistic indicator of classification is the *f1-score* [25]. Composed as a harmonic mean of *precision* and *recall*, the score is computed as follows [25]:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Precision can be understood as the probability of a true positive instance being returned, and is a "positive predictive value" [25]. Recall is a measure of sensitivity, and represents the probability of a positive instance will be returned [25]. F1 scores belong to the broader class of F-scores, where recall is weighted by some value β [25]:

$$F_\beta = (1 + \beta^2)\frac{pr}{r + \beta^2 p}$$

Given the nature of this study, one might ascribe higher priority to detecting cases of suicidality, regardless of any expense of more numerous false alarms on cases of depression. With this in mind, the F2 scores of the classifiers for Depression vs. SuicideWatch are as follows:

Table 3. Optimized f2 Scores (Weighted Averages) for Topic Modelling Approach

k = 5 folds	Depression vs. SuicideWatch
Logistic Regression	0.52
SGD	0.47
Stacked Generalization*	0.56

Table 4. Optimized f2 Scores (Weighted Averages) for Word Embedding Approach

k = 5 folds	Depression vs. SuicideWatch
Logistic Regression	0.42
SGD	0.32
Stacked Generalization*	0.38

*Base Predictors of Random Forest and K-Nearest Neighbor, Meta-Predictor of Logistic Regression

The F1 and F2 scores of the classifiers prove insufficiently accurate to be reliable in flagging at risk individuals. It is also important to note that these metrics and the classifiers they describe, are simply how well Reddit posts exemplify sets of features characteristic of a subreddit. It is not yet certain that even if these classifiers were sufficiently accurate, that they would be also sufficiently indicative of the mental state of the authors of the posts.

However, these results may perhaps be descriptive of the relative strengths of each of the combinations of algorithms, that is the pairings of feature selections and model selections, and may provide insight into the nature of the data itself. Topic modeling performed more accurate overall than word embeddings, suggesting that the hidden semantic structure of the documents is relevant. Shing et. al's paper "Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings" utilized topic modeling, however it made neither a comparison of its benefit relative to word embeddings nor use of stacked ensemble learning [7]. The results from this paper may suggest that topic modelling and ensemble learning are potential avenues for refinement in detecting suicidal ideation.

8 LIMITATIONS

8.1 Cross Domain Subproblem

One might argue that the cross domain subproblem can exist between the three datasets studied in this paper, and therefore future work in addressing this could improve the accuracy of classification. A common solution to this problem is labeling a small portion of the target data for assessing its veiled distribution [8]. A Eshwar Chandrasekharan's 2018 study, "The Internets Hidden Rules", found the following:

> Finally, the discovery of widely overlapping norms suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values. We observed that the F1 scores obtained for relatively smaller subreddits (with less than 5000 removed comments) was approximately 68%, despite using state-of-the-art classifiers. This indicates the potential for using cross-community data to augment and improve completely in-domain classifiers. [9]

8.2 Model Selection

The *Fundamental Law of Model Selection* states that "when the goal is to assess a model's predictive performance, goodness-of-fit ought to be discounted by model complexity" [23]. Cross Validation complies with this law in that it attempts to address overfitting and underfitting by separating the training and test sets [23]. However, any significant amount of noisy data that is in the initial dataset that is split into k-folds will still be taken into account during the cross validation, and will therefore be considered in the model, leading to overfitting. This risk is mitigated, however, by using a relatively large and diverse dataset.

9 CONCLUSION

This paper explores a relatively limited scope of the diverse range of potential natural language processing techniques. There exists potential for evaluation of other features and combinations of features, including syntactic features and Latent Semantic Indexing, as well as different pairings with classifiers and ensembles of classifiers. These methodologies may have varying results across other domains, both internally on Reddit, and externally on other social networks and communities.

The results of this paper suggest that there is considerable variance in predictive strength of selected features and the models used to classify them. Beyond this, there also exists variance in the likelihood that an approach will suggest a prediction that underestimates the severity of a mental health condition.

This study makes the assumption that it is a more beneficial outcome to have classifiers overestimate the severity of a given mental health condition. The rationale of this assumption is derived from Matthew Michael Large's 2018 study "The role of prediction in suicide prevention", which found that "limited sensitivity means that as almost half of the patients who do die by suicide might have been deprived of preventative measures after a lower-risk categorization" [26]. However, the study also suggested that clinicians should have lowered faith in prediction and accept the limits of prevention [26]. There are potential implications resulting from false positive risk assessments. These may include unnecessary restraints on the civil liberties of the subject as a precaution, which consequently consumes societal resources [27]. These implications may also extend to the risk assessing agents, such as reduced credibility in their claims or biased interpretation of risk assessment [27].

REFERENCES

[1] Yari Gvion and Alan Apter MD2. 2012. Suicide and Suicidal Behavior. (December 2012). Retrieved December 5, 2019 from https://doi.org/10.1007/BF03391677

[2] Relias Academy. 2019. Why Suicide Has Become an Epidemic. (December 2019). Retrieved December 5, 2019 from https://reliasacademy.com/rls/store/suicide-epidemic-and-how-to-prevent-suicide

[3] Anon. r/depression - Our most-broken and least-understood rules is "helpers may not invite private contact as a first resort", so we've made a new wiki to explain it. Retrieved December 5, 2019 from https://www.reddit.com/r/depression/comments/doqwow/our_mostbroken_and_leastunderstood_rules_is/

[4] Anon. Suicide Prevention. Retrieved December 5, 2019 from https://www.nimh.nih.gov/health/topics/suicide-prevention/index.shtml

[5] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. (May 2016). Retrieved December 9, 2019 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5659860/

[6] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets. Retrieved December 9, 2019 from https://www.aclweb.org/anthology/P18-3013/

[7] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé Iii, and Philip Resnik. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (2018). DOI:http://dx.doi.org/10.18653/v1/w18-0603

[8] Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution. *Proceedings on Privacy Enhancing Technologies* 2016, 3 (January 2016), 155–171. DOI: http://dx.doi.org/10.1515/popets-2016-0021

[9] Eshwar Chandrasekharan et al. 2018. The Internets Hidden Rules. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (January 2018), 1–25. DOI:http://dx.doi.org/10.1145/3274301

[10] Buffer. [Publish] Character limits for each social network. Retrieved December 5, 2019 from https://faq.buffer.com/article/491-publish-character-limits

[11] Anon. The Python Reddit API Wrapper. Retrieved December 10, 2019 from https://praw.readthedocs.io/en/latest/

[12] Ronen Feldman and James Sanger. Text Mining Preprocessing Techniques. *The Text Mining Handbook*, 57–63. DOI:http://dx.doi.org/10.1017/cbo9780511546914.004

[13] Anon. gensim: topic modelling for humans. Retrieved December 10, 2019 from https://radimrehurek.com/gensim/utils.html

[14] V. Srividhya and R. Anitha. 2010. Evaluating Preprocessing Techniques in Text Categor ization. *International Journal of Computer Science and Application* , 2010 (2010), 49–51.

[15] Natural Language Toolkit. Retrieved December 10, 2019 from https://www.nltk.org/book/ch02.html

[16] Peter F. Brown. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18, 4 (1992), 467–479.

[17] Tomas Mikolov, Bya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems* (December 2013).

[18] Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics - SOMA 10* (2010). DOI:http://dx.doi.org/10.1145/1964858.1964870

[19] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (July 2011), 262–272.

[20] Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (2014). DOI:http://dx.doi.org/10.3115/v1/p14-2050

[21] SciKitLearn sklearn.linear_model.SGDClassifier. Retrieved December 11, 2019 from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier

[22] David Anguita, Alessandro Ghio, Sandro Ridella, and Dario Sterpi. K–Fold Cross Validation for Error Rate Estimate in Support Vector Machines. *Smartware & Data Mining* , 16121.

[23] Quentin Frederik Gronau and Eric-Jan Wagenmakers. 2018. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. (2018). DOI:http://dx.doi.org/10.31234/osf.io/at7cx

[24] Ashley I. Naimi and Laura B. Balzer. 2018. Stacked generalization: an introduction to super learning. *European Journal of Epidemiology* 33, 5 (October 2018), 459–464. DOI:http://dx.doi.org/10.1007/s10654-018-0390-z

[25] Cyril Goutte and Eric Gaussier. 2005. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science Advances in Information Retrieval* (2005), 345–359. DOI:http://dx.doi.org/10.1007/978-3-540-31865-1_25

[26] Matthew Michael Large. 2018. The role of prediction in suicide prevention. *Dialogues Clin Neurosci* , 20 (2018).

[27] Anon. 6.1.2 False Positive and False Negatives in Risk Assessments. Retrieved December 15, 2019 from http://nomsintranet.org.uk/roh/roh/6-best_practice/06_01 _02.htm