Sentiment Analysis On Twitter Data : A Comparative Study

Md Khalid Syffullah
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
md.khalid.syffullah@g.bracu.ac.bd

Khadija-Tul-Kobra
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
khadija.tul.kobra@g.bracu.ac.bd

Sheikh Samiul Huda
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
sheikh.samiul.huda@g.bracu.ac.bd

Syed Abrar Imam
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
syed.abrar.imam@g.bracu.ac.bd

Shovan Kanti Mondal
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
Shovan.kanti.mondal@g.bracu.ac.bd

SHANUMA AFRIN MEGHLA
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
shanuma.afrin.meghla@g.bracu.ac.bd

Sumaiya Ismail
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
sumaiya.ismail@g.bracu.ac.bd

Khairun Nahar
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
khairun.nahar@g.bracu.ac.bd

Humaion Kabir Mehedi
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
Humaion.Kabir.Mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Dept. of Computer Science & Engg, BRAC University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

***Abstract*—Sentiment analysis is actually the process of opinion mining that involves analyzing data from social media, blogging sites, electronic articles, and other sources to determine how people feel about an item, business, organization, or events. Sentiment analysis uses computational linguistics and natural language processing to analyze and extract opinion or sentiment from within text (positive, negative, neutral, etc). Most of the researchers used Naive Bayes, CNN, KNN, LSTM or SVM algorithms in their research work. In this work, we are comparing some researchers proposed methodologies on sentiment analysis for which the source data has been taken from Twitter. One of them, uses K-means(modified) with Naive Bayes and k-nearest neighbors in their research methodology, another researcher group uses KNN classification algorithm, N-Gram for extracting feature and KNN for the classification of sentiment. Analyzing and comparing some recent research works' proposed methods we found a K-means(modified) with Naive Bayes and KNN gives better accuracy.**

***Index Terms*—twitter, Sentiment analysis, Machine learning, Naive Bayes, SVM, KNN, Pattern Recognition**

I. Introduction

The use of social media websites has accelerated in recent years. In today's world, billions of people utilize social media platforms like LinkedIn, Instagram, Tumblr, Facebook, Twitter to share information about their lifestyles and express their views and ideas. People use internet platforms to express their feelings on anything, including items, people, events, and so on. Social networking platforms have two types of advantages. The first is that these platforms allow users to leave feedback and reviews. Another advantage is that businesses can learn more about their products and services in the market. Blog posts, reviews, tweets, comments, status updates, and other forms of social media generate a tremendous amount of sentiment-rich data. Introduction It assists business owners in connecting with their customers for advertising purposes. Natural Language Processing(NLP) is a technique for analyzing authentic text elements. NLP converts text elements into machine-readable formats. Artificial Intelligence makes advantage of the information offered by NLP. It also uses a lot of arithmetic to determine positive and negative thoughts. "Opinion mining" or "emotion AI" are terms used to describe sentiment analysis.

The analysis of sentiment is a method for identifying, quantifying, and studying emotive states and prejudiced data. The most helpful aspect of tweeting is the ability to generate large amounts of sentiment data and analyze it. These statistics are useful in determining people's opinions on a variety of topics. Users find it difficult to assess the content generated. As an outcome, in order to acquire a thorough grasp of distinct user feelings, we need to apply a variety of sentiment analysis

methodologies. Clients may use sentiment analysis to decide whether the information offered about a product is satisfactory before buying it. This data is used by marketers and enterprises for the betterment of their products so that they may be adjusted to their client's demands. The primary goal of textual information extraction methods is to search, process, and interpret accurate data. Even though the reality is dependent on texts, there are some literary contents that express prejudiced property. The main contents of Sentiment Analysis are attitudes, assessments, feelings, views, and emotions. There are numerous challenging opportunities to create new applications. All of this is feasible because of the proliferation of internet sources such as blogs and social media. Sentiment analysis, for example, can be used to forecast which goods would be recommended by an inspection-based recommendation system such as positive or negative attitudes toward those objects. Sentiment analysis encompasses a wide range of activities, including extraction and classification of sentiment, subjective classification, feedback, and opinion spam filtering, to name a few. Its purpose is to discover more about how people feel about products, organizations, subjects, people, and systems. We can argue that, among other social networking sites, Twitter has been a popular micro-blogging site, resulting in a limitless amount of data from which we can obtain opinion mining, as opposed to any other social media. The special symbol hash-tag "'#'" is used to mark topics. Hashtags allow a wider audience to see tweets. During feature selection, all of these valuable tweets are examined alongside the standard textual property. Another key naming convention is the usage of the "@" symbol to indicate target users. When users are mentioned in this way, they are automatically notified. Another important characteristic of tweets is the hashtag.

The following is how the paper is structured: The section "Literature Review" introduces the preferred approaches and reviews. The section "Sentiment classification" describes similar methodologies and categorizes information regarding sentiment analysis. The "Algorithms" section contains the algorithms for several classifiers. It displays the specifics of Twitter sentiment data. "Analyze Proposed Methodology" examines the proposed classifier approach. It also selects suitable algorithms for the system. The outcomes are explained under "result analysis". Finally, "Conclusion" summarizes the entire sentiment analysis as well as future work.

II. Literature Review

The proposed technique is based on an unsupervised learning approach that use huge Twitter corpora to construct embedding by analyzing semantic link that isn't obvious between words dependent on contextual interpretation, as well as statistical aspects of widely used phrases in tweets. Word embedding is used with n-gram and sentiment polarity score characteristics to produce a set of feature from tweets. The feature data set is insert into a deep CNN to develop and classify the sentiment classifier. Applying on our five Twitter datasets, the recommended model outperforms the standard model, with the proposed model outperforming the base model in terms of accuracy and F1-measure [2]. This technique has been presented as a solution to Twitter's challenge of dealing with a wide range of topics. As a result, the categorization algorithm is dynamic. In this method, non-textual features of tweets are also used to train the classification algorithm. The proposed approach can be applied to both static data from various areas as well as dynamic data, which is being streamed in for a specific timeframe. The categorization system divides tweets into three categories: positive, neutral, and negative. In terms of recall, accuracy, and F-score, these three may be combined to create five more class labels: positive, very positive, neutral, negative, and very negative [4]. The approach proposed outperforms the competition. The study offered an add-on technique for improving the SVM classifier's sentiment classification capability, for sentiment analysis and categorization, It is among the most frequently used machine learning methods. The SVM technique is used as the basic classifier, while the Adaboost algorithm is used for Ensemble boosting in this ensemble model [10]. To locate the relationship information between tweets, the proposed algorithm uses structured details connected with tweets such as retweets, tweet followers, and tags inside tweets, and the rest fundamental properties of tweets. This aids in the study of social interaction features. To categorize Twitter data into positive and negative categories, Proposed ensemble model is applied. When compared to the baseline SVM algorithm, the suggested technique outperforms the baseline in terms of accuracy, recall, and F-score. To do sentiment categorization of twitter data, The two simple yet effective ensemble classifiers are Naive Bayes with Mallet's MaxEnt and SentiStrength with Pattern of Textblob to create the two ensemble models [13]. These models solve the challenge of properly training classifiers when there is limited data for training. Both models have demonstrated great sentimental accuracy categorization a total of twelve datasets, proving their efficacy. Both ensemble classifiers are capable of processing enormous volumes of Twitter data and may be run in simultaneously. The suggested methodology examines text at two levels: document and aspect. This method creates a hybrid working model for sentiment analysis and categorization in the future, based on machine learning approaches. The idea of microtext, which has grown in popularity as a result of the expanding usage of Web 2.0 technology, has created issues for standard natural language processing systems that are built to handle well-formed text, according to Ranjan Satapathy, et al research. Microtext normalization aids them in overcoming these obstacles. As a result, the author has presented a phonetics-based method for converting microtext into ordinary English text. The parameter between normalized tweets and tweets standardized by human experts is equal to or better than (8/10=0.8). 31 percent of tweets, according to studies, are spam [8]. Normalizing tweets also gotten better polarity accuracy rate by over 4% [7].

The sentiment analysis approaches studied in this paper are briefly described in this section. NLP may be used to assign polarity, Amazon's Mechanical Turk (AMT) is used to build labeled datasets, psychometric scales are used to determine

mood-based attitudes, and supervised and unsupervised machine learning techniques are used, among other things. The techniques' validation vary a lot as well, from toy examples to a vast amount of labeled data.

Though both publications have merits, I believe the KNN classifier-based research has a better foundation than the other does. As because the study uses a KNN classifier, it has the most scenario and format training.

III. Sentiment classification

The word "sentiment" literally means "emotions". Sentiment analysis is a method of determining the level of public sentiment or opinion on a product or service, as well as a person, such as a politician or a celebrity [1]. Thus, sentiment is a term that describes a topic that is both subjective and objective, as well as a factual or non-factual topic that is neither positive nor negative. Sentiment analysis, often known as opinion mining, is a technique used in data mining. That requires monitoring data from online social networking sites, user reviews, digital media updates, and other sources to learn about people's perceptions of an event, organization, product, brand, person, and so on [2]. The sentiments, or a person's attitude, views, feelings, and opinions, are an important aspect of analyzing a person's behavior. Numerous different people's opinions appear to play a significant role in our decision-making process [3]. Previously, people expressed their views through word of mouth. In recent years, social networking sites have grown in popularity among the general public. People are posting feedbacks and comments on social media networks, and businesses utilize the information to measure how well their products and services perform in the marketplace [2]. Sentimental analysis tries to deduce a user's or author's posture or notion, or an author against an aggressive field or an object. When someone wants to purchase or use a product, for example, the very first step is to read customer reviews and engage in a discussion about it on social media before making a choice. It also helps customers make better selections while purchasing products or using services. Automatic opinion mining have been the most prominent study and research subjects in recent years. However, given the large quantity of data accessible and the structure of the data, creating apps to evaluate it poses a number of challenges [2]. There are three classification levels in sentiment analysis: feature-level classification, document-level classification, and sentence-level classification sentiment analysis [4]. The goal of future-level analysis is to classify people's feelings towards certain entities. This classification also classifies sentiment toward specific features of entities. The first step is identifying the entities and their properties. Diverse people may have different views on the same item's various aspects. The primary purpose of document-level categorization is to classify a viewpoint as positive or negative throughout the content. It treats the entire page as if it were a single entity. Document-level classification also determines if an opinion document reflects a positive or negative mood. It treats the entire article as a single piece of fundamental material. At the sentence level, determining whether a sentence is objective or subjective is the first stage. It will identify if a sentence expresses a negative or favorable view if it is subjective. The sentence-level analysis seeks to classify the emotions formulated in particular sentences. The overall sentiment of each sentence is categorized at the sentence-level classification. Specify if the sentence is subjective or objective as a beginning point. If the statement is arbitrary, the sentence-level determines is it a negative or neutral opinion. Natural expressions of emotion are not necessarily subjective. There is no fundamental distinction between document and sentence level categories because sentences are simply short documents [5]. In general, sentimental analysis can be divided into two ways. One strategy is based on symbolic methods, while the other is based on machine learning. Learning from analogy, discovery, examples, and root learning are some of the learning tactics that are used in the symbolic learning technique. Unsupervised learning and supervised learning are both used in machine learning techniques [4]. For both training and testing, supervised techniques necessitate two sets of annotated data Neural Network, Decision Tree, Support Vector Machine, Naive Bayes, and Maximum Entropy are some of the most often popular machine learning classifiers(supervised). When training data is provided, a supervised technique is viable. In other scenarios, an unsupervised method is taken [3].

IV. Methodology

The research is focused on sentiment classification & detection in the data which is collected from Twitter. Most of the researchers used CNN, Naive Bayes, LSTM, and KNN algorithms in their proposed method.

Hidayat et al. designed a model [7]. To begin, a crawling process is used to collect tweet data from the Twitter API. The Sklearn module of the Python computer language and the Gensim library for the shared phrase-based models and shared memories for data analysis are used to accomplish SVM and Logistic Regression operations.

A model that was proposed by Kundan Reddy proposed [6] incorporates CNN and LSTM algorithms. The LSTM has three gates, all of which are sigmoidal-related. A single node dense layer is used to provide the classification result.

Luo et al. [12] underlined the issues and successful ways of extracting viewpoints from Twitter tweets. Due to spam and drastically varied language, retrieving opinions on Twitter is challenging.

Using The KNN classification method instead of the SVM classification algorithm which had previously been utilized in analytical tools was recommended by Hota and Pathak. Twitter data can be classified into 7 classes using the SVM technique. The goal of this work's KNN classification was to create a classifier to classify Twitter data into 7 categories for analysis of sentiment [2].

Perera et al. analyzed and prepared data for their suggested model. Two models were offered. One employed SVM, while the other used Complement Nave Bayes(CNB) [8].

KNN, Naive Bayes, and Modified K-Means [14] based approach was provided by Bharti and Malhotra. This design consists of four elements: the user interface, log preprocessing, and feature clustering. To identify opinions more correctly, KNN training and testing are utilized. By combining the modified K-means with the Naive Bayes classifier, this system can deal with unrelated data more accurately.

In their suggested model, AlBadani et al. [9] ULMFiT was combined with a classifier, known as SVM, and used in a number of analysis data-sets. RBF kernel also known as Radian basis function, which includes sigmoid, polynomial, and linear functions, was the most basic SVM function.

Naive Bayes technique was utilized in a machine learning model presented by M. Wongkar and A. Angdresey. They must first use a crawler to acquire data from Twitter social media. The tokenization procedure, which involves cleaning the tweet and picking relevant terms, follows. The data is then run through their suggested model [10].

A General Model for sentiment analysis is given below:

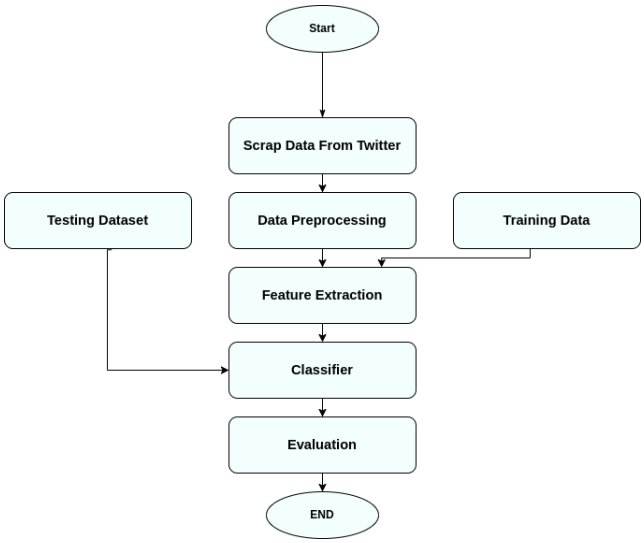

Fig. 1. Proposed Methodology

Almost all of the researchers followed the same process in their research for preparing their data set. There are the steps for their approaches:

1) Data collection for machine learning classifier training and testing.

2) Preparing the data for further analysis by prepossessing it.

3) Natural language processing is used to convert textual input into vector form.

4) Creating training and testing groups using the dataset.

Using training data, the ML Classifier is then trained to predict the polarity of testing data.

A. SVM

SVM classifier, the learnt function divides into positive and negative groups. Let's say, the set of data points is X = x1, x2, x3... and the set of weights is W= w1, w2, w3... . The Maximal-Margin hyperplane divides the data items with weights into two groups. The Maximal-Margin hyperplane can be defined using the Lagrangian formulation [11].

$$D(X^T) = \sum_{i=1}^{l} y_i \alpha_i X_i X^T + b$$

Here, X^T is a tuple, l denotes the support vector's number, and the class label X_i is designate by y_i. α_i and b are real number parameters that are calculated dynamically by the SVM algorithm.

B. KNN

The supervised machine learning method k-nearest neighbor or in short KNN is used to address classification and regression issues [2]. The KNN algorithm maintains all existing data and labels new data points based on characteristics. In KNN classification, the most popular class among its nearest neighbors is assigned the undefined pattern. When two classes have the same average distance from an undefined pattern, the class with the shortest average distance is assigned [14].

C. K-Means

K-Means Clustering is an unsupervised learning approach used in machine learning to tackle clustering issues. The K-means clustering algorithm finds the optimum centroid by calculating centroids and then repeating the procedure. As this algorithm works better with minimal datasets, this algorithm is modified to eliminate several optimum solutions and decrease the use of the cluster-error criteria. To apply this, firstly we need to delete the closest data points by calculating the distance between them and forming a data-point set [14]. Afterwards, the process is repeated again and the next closest data points are removed. This process is repeated until the number of data points reaches a certain threshold [14].

D. Naive Bayes

The Bayes Theorem-based randomized machine learning technique Naive Bayes may be used to tackle a range of classification problems [13]. In this technique, positive and negative sentences are initialized and divided with the number of total sentences. The sentences are then turned into words and classified as positive or negative sentences using Bayes Theorem [13].

E. CNN

CNN, a form of neural network that is used for data processing [9]. The fundamental distinction between CNN and other forms of neural networks is that CNN examines input as a 2D array rather than extracting characteristics [9].

F. LSTM

LSTM networks are Recurrent Neural Network topologies that are meant to recall a limited history of positive values [6]. It is made up of three gates: an input gate that reads in the input, an output gate that writes the output to the following layers, and a forget gate that determines which data to remember and which data to forget [6].

Paper Author	Used Algorithm	Accuracy
Manda et al. [6]	CNN-LSTM	88%
Adisaputra et al. [7]	PV-DM Logistic Regression	87%
Perera et al. [8]	SVM	72.70%
Perera et al. [8]	Complement Naïve Bayes(CNB)	74.96%
AlBadani et al. [9]	LSTM, CNN	79.64%
Wongkar et al. [10]	Naïve Bayes	80.9%
Kharde et al. [11]	SVM	77.73%
Luo et al. [12]	SVM	82.5%
Parikh et al. [13]	Naive Bayes	78.4%
Hota et al. [2]	KNN	86%
Bharti et al. [14]	Modified K-Means + NB + KNN	91%

TABLE I
CLASSIFICATION OF SENTIMENT ANALYSIS APPROACHES AND RESULT.

a) CNN-LSTM: The Long Short-Term Memory Network, or CNN-LSTM, takes embedded words as input and transmits the output to CNN [6]. An algorithm for sentiment representation may be created by merging one layer of CNN with two layers of LSTMs [6]. CNN's convolution layer and pooling layer make it simple to retrieve local features and minimize computing complexity. Aside from that, LSTM is used to learn linguistic syntactic aspects [6].

G. Random Forest

Random Forest is made up of several tree classifiers, which are used to predict the class based on the categorical dependent variable [16]. The input vector is assigned a class by each tree, and the class with the most turns is chosen. The correlation between any two trees in the forest and the strength of any particular tree in the forest define the error rate of this classifier. The trees should be strong and self-contained to limit the rate of errors [17].

In comparison to the proposed systems, K-Means(Modified) with NB and KNN algorithm is more robust and gives more accurate results. The Naive Bayes works exceedingly efficiently for issues that are linearly separable, as well as it also performs moderately well for situations that are non-linearly separable. To tackle clustering problems, the modified K-means algorithm operates best [14].

V. RESULT AND ANALYSIS

According to the table 1, authors applied the different data prepossessing and machine learning approach and complete their work.

Analyzing to their proposed method and output, Bharti, & Malhotra [14] has the highest accuracy rate. The performance is done on the basis of accuracy defined by the following formula.

$$accuracy = \frac{Numberoftweetscorrectlyclassified}{Totalnumberoftweets}$$

The proposed approach using modified K Means algorithm for feature extraction and using KNN and Naïve Bayes combined for classification yields higher accuracy than using N gram for feature extraction and using only Naïve Bayes for classification [14]. The former approach has an accuracy of 91%.

VI. CONCLUSION

We compared different scholars' suggested sentiment analysis approaches in this publication, which is a method for sorting and classifying the feelings indicated by statements from users. In one research, they used the k-nearest neighbors (KNN) algorithm to train the classifier, and they were able to achieve an accuracy of 86% and precision 82% [14]. In another proposed methodology they used various Machine Learning classification algorithms and Natural Language Processing(NLP) on a huge,unbalanced, multi-classed, and real-world dataset providing them 77% accuracy with the Bag-of-Words approach, as well as the Logistic Regression and SVM algorithms. Many researchers reported that Support Vector Machines (SVM) have higher accuracy than other algorithms, but it also has constraints. In another methodology where they have used Naïve Bayes, KNN and k-means(modified) clustering, achieved accuracy of 91% [14]. On a testing dataset of 500 mobile reviews, they achieved accuracy of 91 percent using another methodology that included Naive Bayes, KNN, and k-means(modified) clustering. Furthermore, this algorithm's training and testing times are O (n + V log V) [14], which is quicker than the previous algorithms of machine learning we discussed. KNN and Naive Bayes, in our opinion, are the best approaches for text-based categorization and social interpretation among the presented methods.

REFERENCES

[1] Wongkar, M., & Angdresey, A. (2019, October). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. In 2019 Fourth International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

[2] Hota, S., & Pathak, S. (2018). KNN classifier based approach for multi-class sentiment analysis of twitter data. Int. J. Eng. Technol, 7(3), 1372-1375.

[3] Priya, R. C. M., & Sathiaseelan, J. G. R. (2017, February). An Explorative Study on Sentiment Analysis. In 2017 World Congress on Computing and Communication Technologies (WCCCT) (pp. 140-142). IEEE.

[4] Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017, March). Comparative study of machine learning techniques in sentimental analysis. In 2017 International conference on inventive communication and computational technologies (ICICCT) (pp. 216-221). IEEE.

[5] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

[6] Manda, K. R. (2019). Sentiment Analysis of Twitter Data Using Machine Learning and Deep Learning Methods.

[7] Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Computer Science, 197, 660-667.

[8] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In International conference on advances in ICT for emerging regions (ICTer2012) (pp. 182-188). IEEE.

[9] AlBadani, B., Shi, R., & Dong, J. (2022). A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Applied System Innovation, 5(1), 13.
[10] Wongkar, M., & Angdresey, A. (2019, October). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. In 2019 Fourth International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.
[11] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.
[12] Luo, Z., Osborne, M., & Wang, T. (2015). An effective approach to tweets opinion retrieval. World Wide Web, 18(3), 545-566.
[13] Parikh, R., & Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report, 118.
[14] Bharti, & Malhotra, S. (2016).SENTIMENT ANALYSIS ON TWITTER DATA. pg. 601-609.
[15] K.P.Bennet and C.Campbell. "Support Vector Machines: Hype or Hallelujah?" in Proc.SIGKDD Explorations, 2000, vol. 2, no. 2, pp 1-13.
[16] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[17] Breiman, L. (2015). Random forests leo breiman and adele cutler. Random Forests-Classification Description, 106.