Project Sentiment Analysis

1. INTRODUCTION

WHAT IS SENTIMENT ANALYSIS

Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer’s feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract .

By nature, humans are said to be subjective and often biased with preconceived notions. Modern human societies are largely influenced by opinions in almost all spheres and domain of human civilization. Sentiment Analysis of opinions can, therefore, be viewed as an integral tool to analyse the mood and prevalent disposition of any sample group of individuals concerning any particular product, service, event or topic expressed in text form and published on social media platforms, blog posts, comments, web reviews, etc. With the proliferation of Internet and smart phones even in rural regions in underdeveloped countries around the world, opinions and reviews can be posted with ease. From the point of view of mining such data and opinionated text material, the challenge of examining all posts and reviews and assimilating such into meaningful orientation can be quite considerable. The idea of Sentiment Analysis (SA) has been, therefore, to table a summary of opinions segregated into positive, neutral and negative reviews based on analyses of texts posted by users in assorted digital platforms on the Internet

**Approaches :**

We first use jupyter or kaggle for the given problem statement then we read the csv Test data set file and show the values later we read the training data set csv with which we trained our model . we took text and label as x test and later the label dataset fo y and plot the graph using matplotlib.

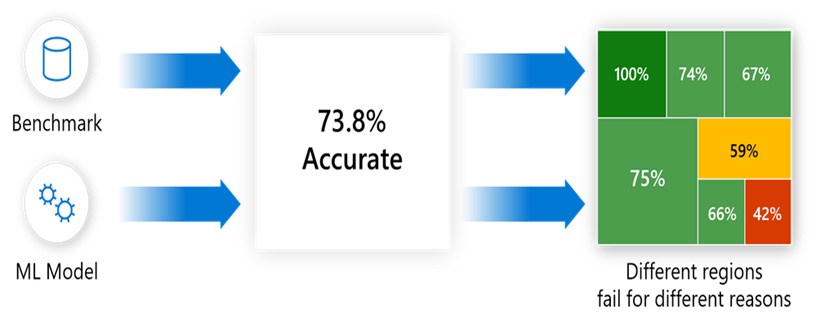
Next, we configure instantiate a TfidfVectorizer model from pretrained checkpoint and coverts the data in vector as it is easy to read .

We train the model with the given data using Support Vector Machines Classifiers and deploy the model .

**What is Error Analysis?**

Error analysis is the process to isolate, observe and diagnose erroneous ML predictions thereby helping understand pockets of high and low performance of the model. When it is said that “the model accuracy is 90%” it might not be uniform across subgroups of data and there might be some input conditions which the model fails more. So, it is the next step from aggregate metrics to a more in-depth review of model errors for improvement.

An example might be that a dog detection image recognition model might be doing better for dogs in an outdoor setting but not so good in low-lit indoor settings. Of course, this might be due to skewed datasets and error analysis helps identify if such cases impact model performance. The below illustration provides a view of how moving from aggregate to group-wise errors provides a better picture of model performance.



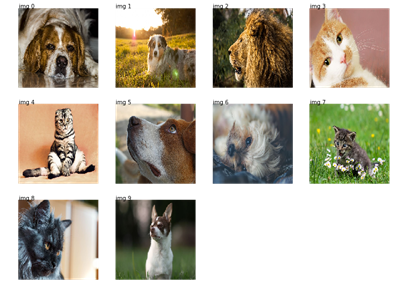
## Error Identification and Diagnosis

This helps know how errors are distributed across key hypotheses or key features/classes/cohorts of the dataset. For example, in a loan approval model used by a bank, it might be that the model is giving more errors on the individuals who are younger and have a low monthly average balance with the bank.

A. How to do this (manually) especially in case your data is image, voice, or text where you might not have apparent features.

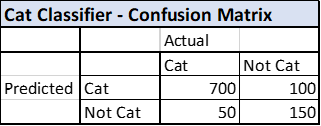
Let us take another example – a favorite one – A cat-classifier! Suppose we train a model on 4000 images (Some having cats) in different settings and test using 1000 images. We find that the model accuracy is 85% (meaning 850 of the 1000 images were predicted correctly!) and 150 images were wrongly tagged.

### Sample of Images with Erroneous Classification in Cat-classifier



Source: http://www.xavierdupre.fr/app/mlinsights/helpsphinx/notebooks/search\_images\_keras.html

On making a confusion matrix it looks like below:



Source: Created by the author

But it does not help understand what is being misclassified, so we isolate the 70 and 80 images respectively, so we look at the images that have been misclassified and create some hypotheses/observations on grouping these errors.

On reviewing all examples quickly, we identify that the following seems to be common across all erroneous cases:

1. White cats not being correctly identified

2. Dogs like cats being identified as cats

3. Dark background leads to misclassification and so on

The hypothesis or commonalities observation for all erroneous use cases is followed by creating a table in Excel or a similar tool to map the exact distribution of the errors.

**CONCLUSIONS**

This report discusses in details the various approaches to Sentiment Analysis, mainly Machine Learning and Cognitive approaches. It provides a detailed view of the different applications and potential challenges of Sentiment Analysis that makes it a difficult task.

41 We have seen the applications of machine learning techniques like Naïve Bayes, Maximum Entropy, Support Vector Machines and Voted Perceptrons in SA and their potential drawbacks. As all of these are bag-of-words model, they do not capture context and do not analyze the discourse which is absolutely essential for SA. We have also seen the use of Subsequence Kernels in Voted Perceptrons that is somewhat successful to capture context as a result of which it achieves a high accuracy. Also it achieves the difficult task of performing prediction over a continuum even though trained only on the extreme reviews. Thus machine learning models with a proper kernel that can capture the context will play an important role in SA.