Twitter Sentiment Analysis, Project Proposal

Problem Statement:
The problem statement is to "predict" the sentiment (positive, negative or neutral) from the content of the tweets ("tweet" column). The ground truth (i.e., true class labels) is determined from the "class" columns. This is a three-class classification problem.

Team Members

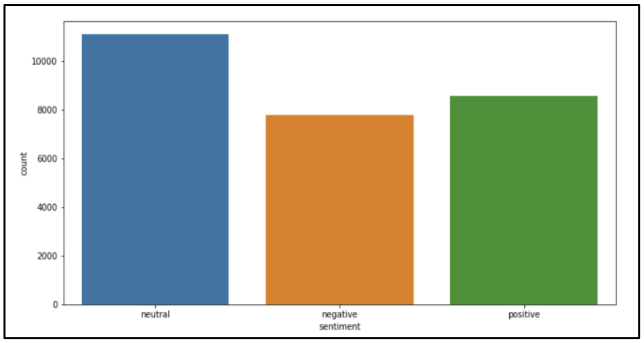
	Team Members	Deliverables
1	Muhammad Abrar Tariq	Exploratory Data Analysis, ML Model
2	Ajay Sagar	Evaluation, ML Model
3	Chinmay Tarwate	Data Cleaning & Preprocessing, ML Model
4	Varun Bhalla	Data Cleaning & Preprocessing, Baseline, ML Model

* We plan to use dictionary based method as our baseline and compare it with Support Vector Machines, Logistic Regression etc.

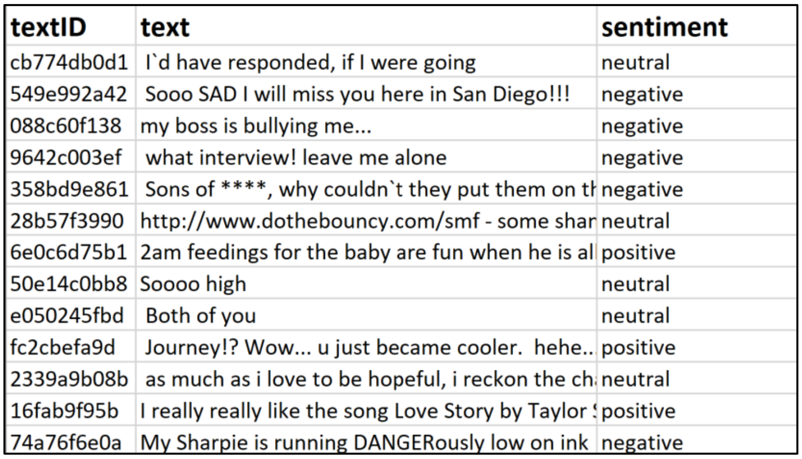
About the Dataset
We use a dataset from Kaggle – "Twitter Sentiment Analysis". The dataset consists of tweets extracted from twitter and is pre-divided into train and test sets. This is a three class-classification problem and the dependent variable is either "Positive", "Negative" or "Neutral". The training dataset has about 27k+ rows and the testing dataset has about 3k+ rows. The dataset looks almost balanced, and the distribution is as follows.

Sentiment	Rows
Neutral	11117
Positive	8582
Negative	7781

Datasets	Rows
Training	27481
Testing	3534



Data Distribution



textID	text	sentiment
cb774db0d1	I`d have responded, if I were going	neutral
549e992a42	Sooo SAD I will miss you here in San Diego!!!	negative
088c60f138	my boss is bullying me...	negative
9642c003ef	what interview! leave me alone	negative
358bd9e861	Sons of ****, why couldn`t they put them on th	negative
28b57f3990	http://www.dothebouncy.com/smf - some shar	neutral
6e0c6d75b1	2am feedings for the baby are fun when he is al	positive
50e14c0bb8	Soooo high	neutral
e050245fbd	Both of you	neutral
fc2cbefa9d	Journey!? Wow... u just became cooler. hehe..	positive
2339a9b08b	as much as i love to be hopeful, i reckon the ch	neutral
16fab9f95b	I really really like the song Love Story by Taylor S	positive
74a76f6e0a	My Sharpie is running DANGERously low on ink	negative

A snapshot of RAW datasets

Machine Learning Techniques: Plan Outline

The problem proceeds in three stages:

1. **Text Cleaning & Processing**: We will clean up the raw tweet text using the various functions like removing punctuations, lemmatizing etc. and convert them into tokenized data
2. **Exploratory Analysis & Feature construction**: We create bag-of-words feature vectors and training labels from the processed text of tweets and the class columns respectively.
3. **Classification & Evaluation**: We use a dictionary-based method to create a baseline model. After the features are created, we use them to learn four different machine learning models which can classify them based on their sentiment and evaluate our models against the baseline.

Project Pipeline

A. Text Cleaning & Processing

We will create a function which processes raw text and convert it into tokens. All tokens are processed as follows.

1. The tokens will be converted to lower case and appear in the same order as the original text.
2. The tokens will be converted in their lemmatized form (if possible)
3. All punctuations are to be removed. (string.punctuation)
4. Words are separated when hyphen is encountered.
5. All URLs are removed

B. Exploratory Analysis & Feature Construction

1. After tokenization, we will do the exploratory data analysis.
2. From the tokenized data we will use the TF-IDF feature vector methodology to create feature vectors.
3. We will create a sparse matrix for each of the tweets.
4. We will also use grid search to find out the best way to create feature vectors. Since we study tokens as individual units and their relationships to sentiments, we will also study and examine which words tend to follow others immediately or that co-occur within the same document (n-gram model)

C. Classification & Evaluation

1. We will create a baseline classifier which uses dictionary-based methods to classify the tweets. Dictionary based methods are unsupervised techniques which predicted the sentiment of the text using knowledgebases, ontologies, databases, and lexicons specially curated and prepared just for sentiment analysis. They work well when the text is subjective and has various emotions like “sad”, “bad”, “good”, “happy” etc. Most of these lexicons have a list of positive and negative polar words and some score associated with them. Scores are assigned to the text for which we want to compute the sentiment using this method. After aggregating these scores, we get the final sentiment. (We use AFINN lexicon dictionary for our classifier)
2. After we create our features, we learn four different models (eg. Support Vector Machines, Logistic Regression, Random Forest and XGBoost) and compare it against our baseline classifier on the accuracy metric. We plan to use cross-validation and grid search in order to evaluate our models better.