Name: Rohit Garla Net ID: rg3365
Name: Mengrui Zhang Net ID: mz2258

Sentiment Analysis Project

Problem & Background:

Twitter is a popular microblogging website. Each tweet is 140 characters in length. Tweets are frequently used to express a tweeter's emotion on a particular subject. The challenge is to gather all such relevant data, detect and summarize the overall sentiment on a topic. The problem in sentiment analysis is classifying the polarity of a given text at the document sentence or feature level.
Our Objective is to build a classifying model and compare it with another model, to draw relevant insights on the data. Obtained. We have performed Sentiment analysis, to determine the attitude of mass is positive, negative or neutral towards the subject of interest.
Our target is to use the contextual information of the tweets in order to categorize it to a given class (positive/neutral/negative). Following the standard bag-of-words framework that is commonly used in natural language processing and information retrieval. Due to the minimum assumptions that the Maximum Entropy classifier makes, we regularly use it when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions.

Moreover, Maximum Entropy classifier is used when we can't assume the conditional independence of the features. This is particularly true in Text Classification problems where our features are usually words which obviously are not independent

An aspect of social media such as twitter messages is that it includes rich structured information about the individuals involved in the communication. It can lead to more accurate tools for extracting semantic information.

Data:

The Data is collected by calling the twitter api on topic "Avengers Infinity War". We have used nltk.twitter package to download 1000 tweets for our model. We have downloaded two dictionary, which were used for classifying the tweets into positive, negative and neutral from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

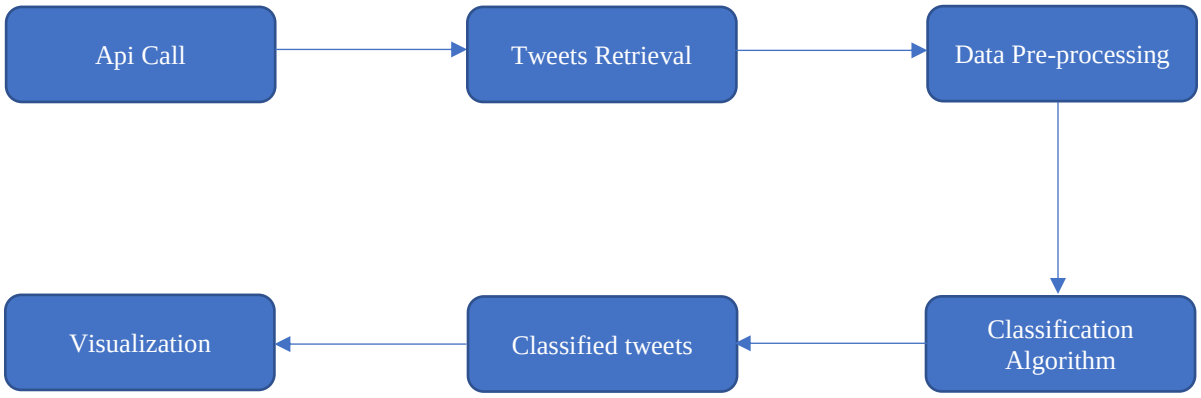
Data and Model Analysis:

A quantitative analysis of data has helped us answer the following questions:

1. What is mass opinion of the movie 'Infinity War'?
2. Which tweets were the most retweeted ones?
3. Which model is more suited for tweet analysis?

Name: Rohit Garla Net ID: rg3365
Name: Mengrui Zhang Net ID: mz2258

Proposed System:



Problem Statement:

Movie reviews are a crucial aspect of sensitizing an audience towards a movie. They provide people with multiple perspectives and allow people to decide whether to watch a movie. Movie reviews are far more personalized than statistical ratings or scores. They embody the opinion of the reviewer and are highly subjective in nature. A crucial characteristic of the reviews is their sentiment, or overall opinion towards the subject matter. As such, sentiment analysis is performed to systematically identify, extract, quantify, and study the reviews.

The problem statement involves:

1. Building a classifier for polarity detection of movie reviews.
2. Training and testing the classifier using a huge set of positive and negative reviews.
3. Performing sentiment analysis and classification - Uncovering the attitude of the author on a particular topic from the written text; alternatively known as “opinion mining” and “subjectivity detection”.
4. Using natural language processing and machine learning techniques to find statistical and/or linguistic patterns in the text that reveal attitudes.

Type of Model:

Maximum Entropy:

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naive Bayes classifier that we will discuss in the next section, the Max Entropy does not assume that the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

Name: Rohit Garla Net ID: rg3365
Name: Mengrui Zhang Net ID: mz2258

Naïve Bayes:

Learning in a naïve Bayes classifier is a simple matter of counting up the number of co-occurrences of features and classes. With the Naive Bayes model, we do not take only a small set of positive and negative words into account, but all words the NB Classifier was trained with, i.e. all words present in the training set.

Problem in Scope of Class:

Data Pre-processing:

During the course of this project, we cleaned the data, created many new features and Dataframes which were then used in further analysis and classification.

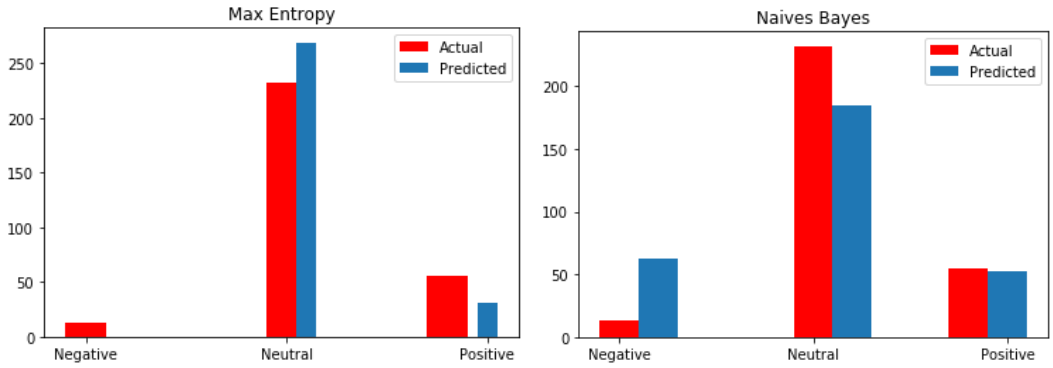
The Naïve Bayes and KNN classifiers are used, along with some pre-processing steps. The pre-processing steps are:

1. Removal of "stop words": Stop words are certain words which don't make a difference to the result. These words are removed from the dataset so as to just concentrate on words that matter in predicting the polarity of the movie review. A very common English stop word list provided on (http://www.ranks.nl/resources/stopwords.html) is referred to for the stop word list.
2. Stemming: Stemming is the process for reducing inflected/derived words to their stem, base or root form—generally a written word form. The stem is such that related words map to the same stem, even if this stem is not a valid word. Stemming is done via Porter Stemming algorithm, and results in: car, cars, car's, cars' => car
3. Removal of any word of two characters or less: Such words are typically punctuations and are thus removed.

Data Visualization:

We visualized many values and tried to understand the underlying pattern in each of them. The Visualization helped us to get more information from the data and guided us in making decisions that were implemented later on in the project.

Name: Rohit Garla Net ID: rg3365
Name: Mengrui Zhang Net ID: mz2258



Assumptions and Limitations:

- The Max Entropy requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model.
- When we have dependent features MaxEnt classifiers produces high results in performance, compared with NB. On the other hand, MaxEnt produces poor results when we have independent features where NB scores high performance.

Evaluation Metrics:

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while **recall** (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.
A measure that combines precision and recall is the harmonic mean of precision and recall

Maximum Entropy:

Accuracy: 87.66 %

Precision Positive: 1.0
Recall Positive: 0.56
F – Positive: 0.72

Naive Bayesian:

Accuracy: 77 %
Precision Positive: 0.77
Recall Positive: 0.70
F – Positive: 0.72

Precision Negative: 0.17
Recall Negative: 0.84
F – Negative: 0.28

From above result, we can find that Maximum Entropy model is more passive than Naive Bayesian model, therefore, Maximum Entropy model realized a higher precision, but a lower recall. It's makes sense, since Maximum Entropy is "safer" model than others, and It tends to classify more record into neutral label which is the majority of our dataset. The tradeoff between precision and recall here, makes two models has the same F1 score which is 0.72. Then we can conclude that two models are both reasonable and meaningful. However, the basic measure of accuracy may give us a hint that Maximum Entropy model may perform better than Naïve Bayesian model in this problem.

Future Work for improving Accuracy and Conclusion:

Sentiment analysis play vital role to make decision like movie review domain. Due to its tremendous value for practical applications, there has been an explosive growth of both research in academia and applications in the industry. In this study, an experiment is conducted on Movie Review dataset. The Naive Bayes classifier and Maximum Entropy classifier is used to train dataset. Accuracy of sentiment analysis is increased by proposed system from dependence and independence assumptions among features. In future, apply this work on clustering domain for movie review dataset for opinion mining applications where the cluster-based features are used to address the problem of scarcity of opinion annotated data in a language. Maximum Entropy will not be hurt by strong independence assumptions, as would naïve Bayes with these features. In Future, we can improve the accuracy with expanded feature classes.