Twitter Sentiment Analysis

Abstract

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analysis of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams. We have also discussed general challenges and applications of Sentiment Analysis on Twitter**.**

Introduction

Nowadays, the age of Internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media,etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision. The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used.

Sentiment analysis (SA)tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about

their products or services in such a way that it can be offered as per the user‟s requirements.

Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but,there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by considering considerations such as positive or negative opinions about those items by making use of SA.

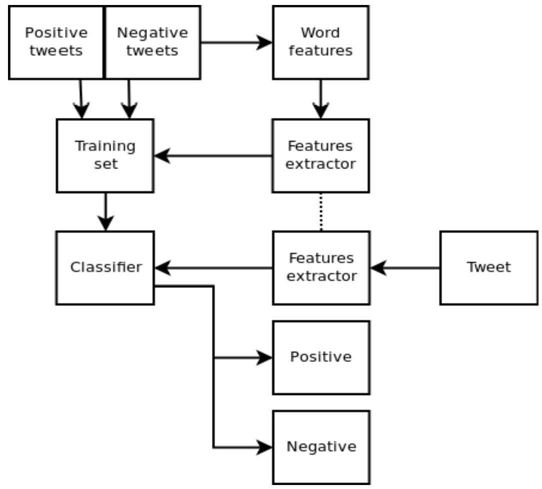


Fig.1. Sentiment Analysis Architecture

This work is structured as follows: we first discuss our procedure for building machine learning model and explain the basics of machine learning technique used for text classification problems (Section 2). Then, we summarize the findings from the literature review conducted to understand the research field and to identify gaps in knowledge. Once these gaps are identified, we lay out the details of our system and our motivation in choosing the features for this system. We then run experiments on this system to compare the performance of our system with other published work. Also, the effectiveness of each feature added to the system is highlighted. Lastly, our research contributions are summarized with future steps needed to be taken (Section 3).

The second section discusses our procedure for building interactive visualization tool that uses sentiment prediction model. We summarize the findings from the literature review conducted to understand what metrics and features have worked well and how can they be adopted to sentiment analysis visualization. We also describe other related public tools that perform the similar task. After highlighting past work, we describe our approach and system architecture for developing the visualization tool. Next, we present the screenshots of the application describing the motivation for each of user view and element. Lastly, we summarize our contributions and future steps needed.

Machine Learning Background

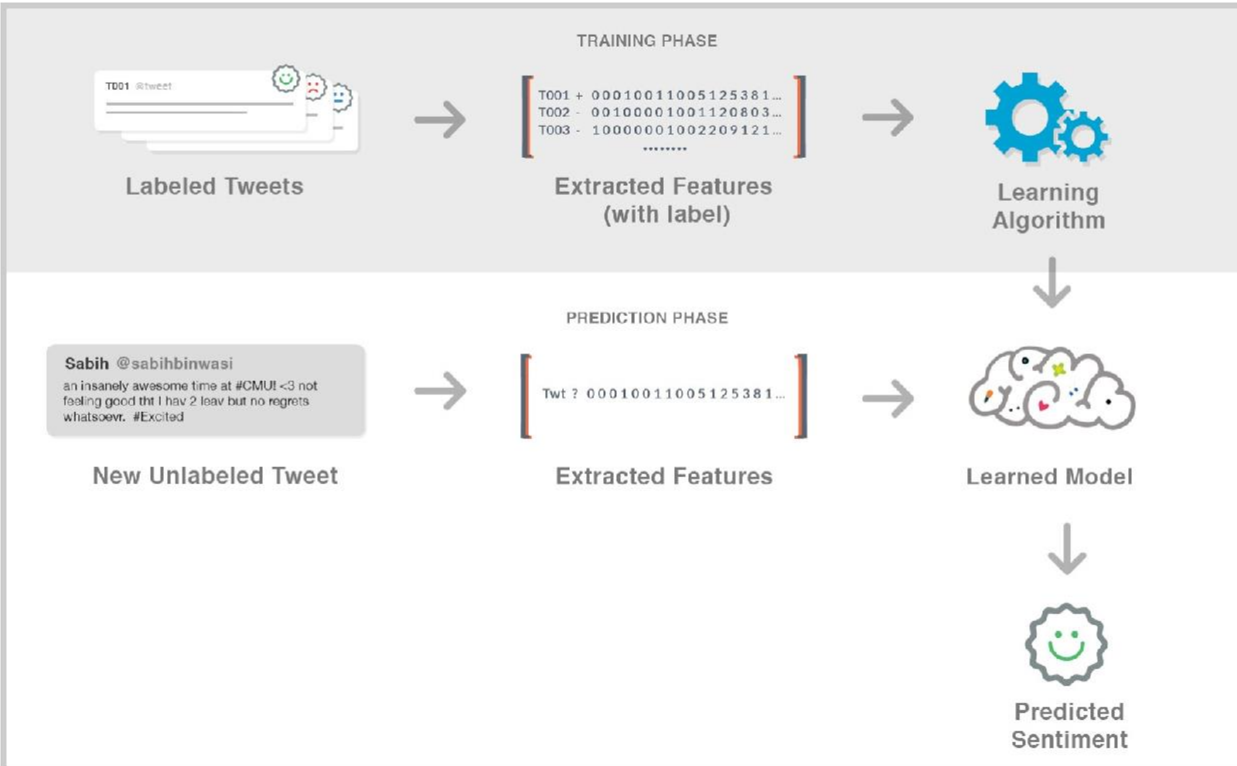


Figure 1: Overview of Supervised Sentiment Classification of Tweets

Before understanding research conducted for Twitter sentiment analysis, we need to describe standard procedure to tackle this problem. Supervised Text Classification, a machine learning approach where a class predictor is inferred from labeled training data, is a standard approach for sentiment classification of tweets. The overview of this approach, adapted to sentiment classification of a tweet, is illustrated in figure 1. First, a dataset of labeled tweets are compiled. Each tweet in this set is labeled positive, negative or neutral by a human annotation based on the sentiment the annotator(s) concluded after analyzing the tweet

Then, a feature extractor generates a feature vector for each labeled tweet where the values of this vector should characterize the sentiment. Once feature vectors are extracted for each tweet in the labeled dataset, they are fed to classification algorithm that attempts to find relations between each value (called feature) in the vector and the labeled sentiment. Popular classification algorithms used for this task are SVM (support vector machines), Naive Bayes and Maximum Entropy method.

Preprocessing

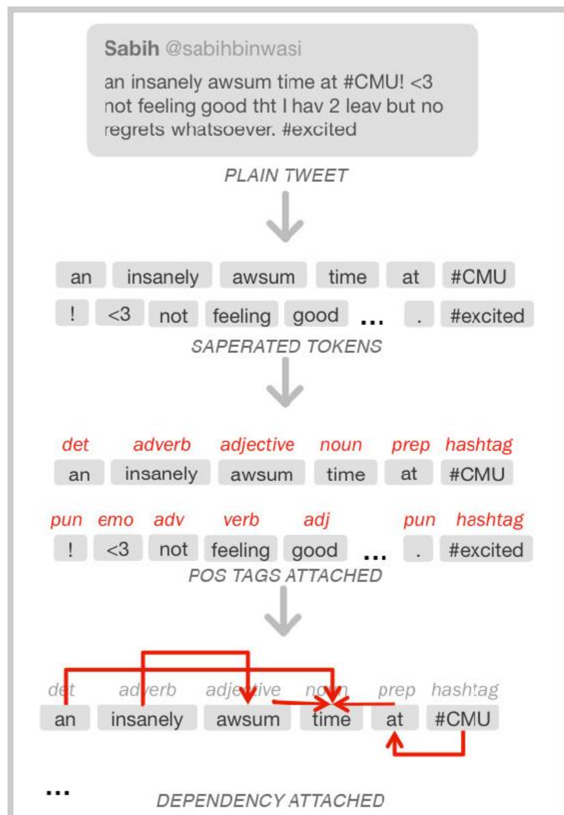
A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

- Remove all URLs (e.g., www.xyz.com), hash tags (e.g. #topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled

- Replace all the emoticons with their sentiment.
- Remove all punctuations, symbols, numbers
- Remove Stop Words
- Expand Acronyms (we can use a acronym dictionary)
- Remove Non-English Tweets

Before the feature extractor can use the tweet to build feature vector, the tweet text goes through preprocessing step where the following steps are taken. These steps convert plain text of the tweet into processable elements with more information added that can be utilized by feature extractor. For all these steps, third-party tools were used that were specialized to handle unique nature of tweet text. These tools were developed by a research group at Carnegie Mellon University (Gimpel et al., 2011; Kong et al., 2014). All three preprocessing are illustrated in Figure 2.

Step 1: Tokenization



Tokenization is the process of converting text as a string into processable elements called tokens. In the context of a tweet, these elements can be words, emoticons, url links, hashtags, or punctuations. As seen in Figure 2, "an insanely awsum…." text was broken into "an", "insanely", "awsum".

These elements are often separated by spaces. However, punctuation ending the sentence like exclamation marks or full-stop are often not separated by a space. On the other hand, hashtags with "#" preceding the tag needs to be retained since a word as a hashtag may have different sentiment value than a word used regularly in the text. Therefore, Twitter-specific Tokenizer [8] is used to extract tokens.

Step 2: Parts of Speech Tags

Parts of Speech (POS) tags are characteristics of a word in a sentence based on grammatical categories of words of a language. This information is essential for sentiment analysis as words may have different sentiment value depending on their POS tag. For example, word like "good" as a noun contains no sentiment whereas "good" as an adjective reflect positive sentiment.

As seen in Figure 2, each token extracted in the last step is assigned a POS tag. Like in Step 1, we use Twitter-specific POS tagger (Gimpel et al., 2011) that can handle out-of-vocabulary words (OOV) and Twitter related tags like hashtags and emoticons. The accuracy of this POS tagger is 93%.

Step 3: Dependency Parsing

For our purposes, dependency parsing is extracting the relationship between words in a sentence. This can be useful in identifying relationship between “not” and “good” in phrases like “not really good” where the relationship is not always with the adjacent word. The dependency parsing tool we use (Kong et al., 2014) gave parent-child relationship between tokens in a tweet as seen in Figure 2. The accuracy of the dependency parser is 80% on Twitter data.

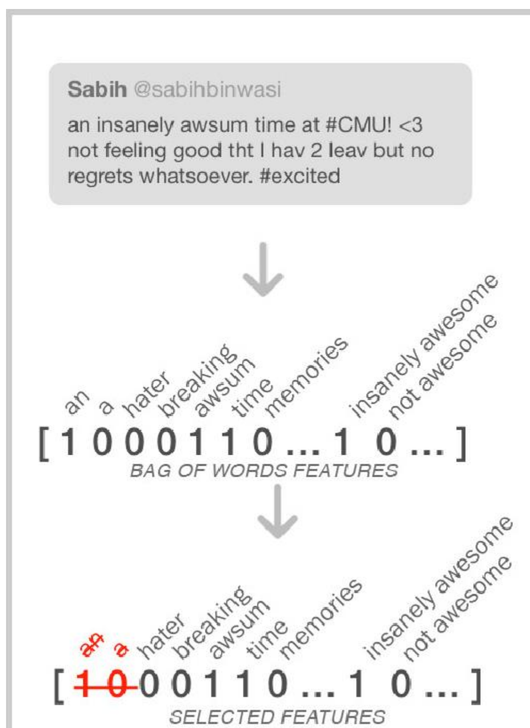
Feature Extraction

Feature extraction is the process of building a feature vector from a given tweet. Each entry in a feature vector is an integer that has a contribution on attributing a sentiment class to a tweet. This contribution can vary from strong, where the value of a feature entry heavily influences the true sentiment class; to negligible, where there is no relationship between feature value and sentiment class. It is often the job of classification algorithm to identify the dependency strength between features and classes, making use of strong correlated features and avoiding the use of ‘noisy features.

During this work, we followed incremental feature selection approach. After adding each set of features, we conducted experiments to compare the performance of the system before and after the feature addition. If the feature set significantly improved the performance of the system, it remained in the feature vector; otherwise, it was removed.
Performance was evaluated using a standard cross validation method. The training dataset was pseudo randomly divided into ten non-overlapping subsets preserving the class distribution. The accuracy of the system was evaluated using nine folds as the training set and one-fold as the evaluation set. This process occurs ten times, each time selecting a different fold as evaluation set. Once all ten accuracy scores are collected, an average is calculated that we use as a performance metric.

1.1.1 Bag of Words Feature Set

Bag of Words (unigrams) is a set of features where the frequency of tokens (or in our case, presence of a token) is indicated in a feature vector. From our study of past work, this feature set was unanimously chosen by researchers to be included in the feature vector. An entry in the feature vector is assigned to each unique token found in the labeled training set. If the respective token occurs in a tweet, it is assigned a binary value of 1 otherwise it is 0. Note that the grammar structure or ordering of token sequence is not preserved. Instead, only the independent presence of a token preserved. As seen in Figure 3, token “awsum” occurs in a given tweet and hence that column gets a value of 1. On the other hand, the word “hater” does not occur in a tweet and therefore has a value of 0. Note that, “hater” would have occurred in some tweet instance in the training dataset since we have a column for it in a feature vector.



The effectiveness of using bag of words for sentiment analysis have been reported by various publications (Pang et al., 2002; Mohammad et al., 2013; Bin Wasi et al., 2014). It was also found that indicating only presence of a word yields higher performance than indicating the frequency of a word. The underlying cause for this behavior may be that the sentiment class is not usually varied if certain words occur more than once in the text.

Training

Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data.

Machine Learning Approaches

Machine learning based approach uses classification technique to classify text into classes. There are mainly two types of machine learning techniques

Unsupervised learning:

It does not consist of a category, and they do not provide with the correct targets at all and therefore rely on clustering.

Supervised learning:

the model during the process. These labeled datasets are trained to get meaningful outputs when encountered during decision- making.

The success of both this learning methods is mainly depending on the selection and extraction of the specific set of features used to detect sentiment.

The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In a machine learning technique, two sets of data are needed:

1. Training Set
2. Test Set.

A number of machine learning techniques have been formulated to classify the tweets into classes. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis.

Machine learning starts with collecting training dataset. Nextly we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is to select feature. They can tell us how documents are represented.

The most commonly used features in sentiment classification are

- Part of speech information
- Negations
- Opinion words and phrases

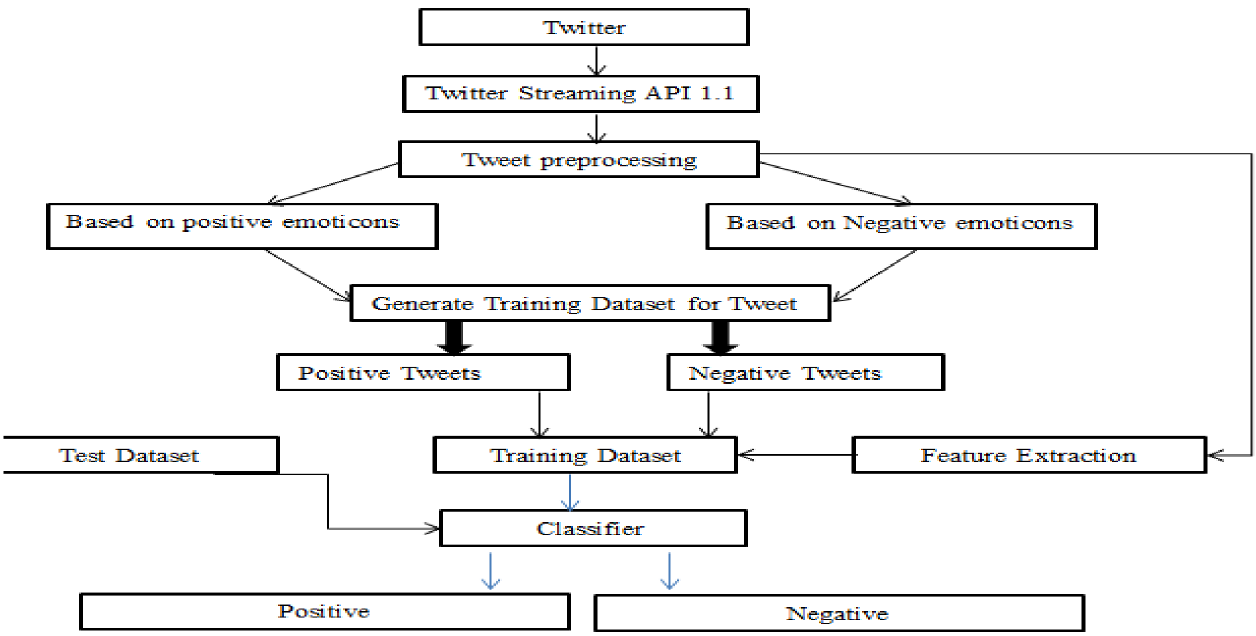


Fig.2 Sentiment Classification Based on Emoticons

With respect to supervised techniques, support vector machines (SVM), Naive Bayes, Maximum Entropy are some of the most common techniques used.

Whereas semi-supervised and unsupervised techniques are proposed when it is not possible to have an initial set of labeled documents/opinions to classify the rest of items

LEVELS OF SENTIMENT ANALYSIS

Tasks described in the previous section can be done at several levels of granularity.

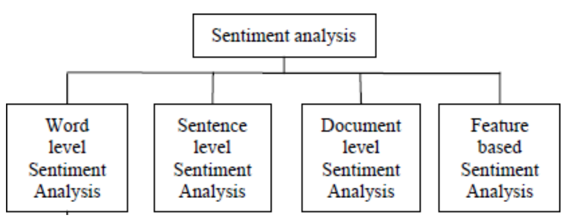


Fig.5 Levels of Sentiment Analysis

EVALUATION OF SENTIMENT CLASSIFICATION

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP) Recall = TP/(TP+FN)

F1 = (2×Precision×Recall)/(Precision+Recall)

In which TP, FN, FP, and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances, as defined in the table 1.

Table 3. Confusion Matrix

	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

RESULTS

We used the twitter dataset publicly made available by Stanford university. Analyses was done on this labeled dataset using various feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.

Dataset Description:

Train Data	45000
Negative	23514
Positive	21486

Test Data	44832
Negative	22606
Positive	22226

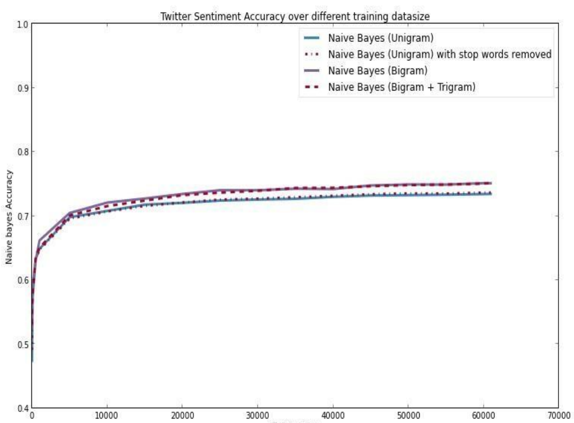


Fig.6 Graph Representing Different results obtained for Naïve Bayes Algorithm.

Table 8. Accuracy of Naïve Bayes Algorithm

Algorithm	Accuracy
Naïve Bayes (unigram)	74.56
Naïve Bayes (bigram)	76.44
Naïve Bayes (trigram)	75.41

Table 10. Summary for Accuracy of SVM Algorithm

Algorithm	Accuracy
SVM with unigram	76.68
SVM with bigram	77.73

Table 11. Summary Of Results For Unigram

Method	Accuracy (Unigram)
Baseline	73.65
Naïve Bayes	74.56
SVM	76.68
Maximum Entropy	74.93

As the table shows, when the processing, analysis was done on the bigger dataset, the accuracy scaled upto a great extent. Naive Bayes baseline scaled upto 76.44 and SVM scaled upto 77. 73percent.The best result tested thus far, was obtained when SVM was used on a feature set of a combination of Unigram, Bigram with stop words removal, gave an accuracy of
77.73. Maxent also performed well and gave an accuracy of 74.93when stop words was removed.

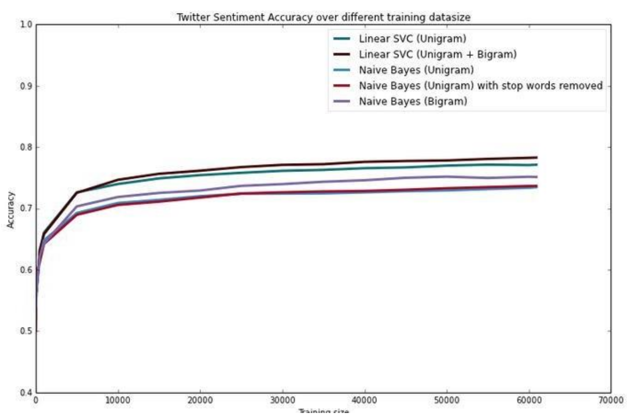


Fig.7 Graph Representing Different results obtained for Naïve Bayes Algorithm And Linear SVC (SVM).

CHALLENGES IN SENTIMENT ANALYSIS

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

- **Domain dependence:**

The same sentence or phrase can have different meanings in different domains. For Example, the word „unpredictable" is positive in the domain of movies, dramas, etc, but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.

- **Sarcasm Detection:**

Sarcastic sentences express negative opinion about a target using positive words in unique way.

- **Building a classifier for subjective vs. objective tweets.**

Current research work focuses mostly on classifying positive vs. negative correctly. There is need to look at classifying tweets with sentiment vs. no sentiment closely.

- **Handling comparisons.**

Bag of words model doesn't handle comparisons very well.

- **Applying sentiment analysis to Facebook messages.**

There has been less work on sentiment analysis on Facebook data mainly due to various restrictions by Facebook graph API and security policies in accessing data.

APPLICATIONS OF SENTIMENT ANALYSIS

1. *Applications that use Reviews from Websites:*

Today Internet has a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus, there is a need for a sentiment analysis system that can extract sentiments about a particular product or services. It will help us to automate in provision of feedback or rating for the given product, item, etc. This would serve the needs of both the users and the vendors.

2. *Applications as a Sub-component Technology*

A sentiment predictor system can be helpful in recommender systems as well. The recommender system will not recommend items that receive a lot of negative feedback or fewer ratings.

In online communication, we come across abusive language and other negative elements. These can be detected simply by identifying a highly negative sentiment and correspondingly taking action against it.

3. *Applications in Business Intelligence*

It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online opinion decides the success or failure of their product. Thus, Sentiment Analysis plays an important role in businesses. Businesses also wish to extract sentiment from the online reviews in order to improve their products and in turn their reputation and help in customer satisfaction.

4. *Applications across Domains:*

Recent research in sociology and other fields like medical, sports have also been benefitted by Sentiment Analysis that show trends in human emotions especially on social media.

5. *Applications In Smart Homes*

Smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a tablet device. Recently there has been lot of research going on Internet of Things (IoT). Sentiment Analysis would also find its way in IoT. Like for example, based on the current sentiment or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment.

Sentiment Analysis can also be used in trend prediction. By tracking public views, important data regarding sales trends and customer satisfaction can be extracted.

CONCLUSION

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning approaches, together with cross domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy. We also studied the effects of various features on classifier. We can conclude that more the cleaner data, more accurate results can be obtained. Use of bigram model provides better sentiment accuracy as compared to other models. We can focus on the study of combining machine learning method into opinion lexicon method to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

Questions and Answers related to Project

Q1. What was the size of the data?
Answer: 1000000 X 6

Q2. What was the data type?
Answer: Int and str

Q3. What techniques were you using for data pre-processing for various data science use cases and visualization?

- Answer: While preparing data for a model, data should be verified using multiple tables or files to ensure data integrity.
- Identifying and removing unnecessary attributes.
- Identifying, filling, or dropping the rows/columns containing missing values based on the requirement.
- Checking null values.
- Identifying, filling or dropping the rows/columns containing missing values based on the requirement.
- Based on the requirement, form clusters of data to avoid an overfitted model.
- Scaling the data so that the difference between the magnitudes of the data points in different columns are not very big.
- Converting the categorical data into numerical data.
- Replacing or combining two or more attributes to generate a new attribute which serves the same purpose.
- Trying out dimensionality reduction techniques like PCA(Principal Component Analysis), which tries to represent the same information but in a space with reduced dimensions.

Q4. What were your roles and responsibilities in the project?
Answer: My responsibilities consisted of gathering the dataset, Cleaning, and getting familiar with the data, model training. Performing various EDA and feature engineering techniques and testing by applying various machine learning models. Also performing hyper-parameter
Tuning for getting optimized results.

Q5. In which area you have contributed the most?
Answer: I contributed the most to data preparation, preprocessing and model training areas. Also, we did a lot of brainstorming for finding and selecting the best algorithms for our use cases. After that, we identified and finalized the best practices for implementation.

Q6. In which technology you are most comfortable?
Answer: I have worked in almost all the fields viz. Machine Learning, Deep Learning, and Natural Language Processing, and I have nearly equivalent knowledge in these fields. But preferred one is loved working in ML.

Q7. In how many projects you have already worked?
Answer: I have worked in various projects e.g., object detection, object classification, object identification, NLP projects, machine learning regression, and classification problems.

Q8. What kind of challenges have you faced during the project?
Answer:

Domain dependence:

The same sentence or phrase can have different meanings in different domains. For Example, the word „unpredictable‟ is positive in the domain of movies, dramas, etc, but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.

- **Sarcasm Detection:**

Sarcastic sentences express negative opinion about a target using positive words in unique way.

- **Building a classifier for subjective vs. objective tweets.**

  Current research work focuses mostly on classifying positive vs. negative correctly. There is need to look at classifying tweets with sentiment vs. no sentiment closely.

- **Handling comparisons.**

Bag of words model doesn't handle comparisons very well.

- **Applying sentiment analysis to Facebook messages.**

There has been less work on sentiment analysis on Facebook data mainly due to various restrictions by Facebook graph API and security policies in accessing data.

Q9. How did you optimize your solution?
- Answer: Train with better data (increase the quality),or do data pre-processing steps more efficiently.
- Increase the quantity of data used for training.
- Hyper- parameter tuning performs regularly.