A Study of Social media sentiment analysis on twitter data using different techniques

Vasava Devyangini Rajeshbhai
Department of Computer Engineering
L. D. College of Engineering, Ahmedabad

Abstract

The explosive growth of social platforms on web including blogs, products review sites, forums, Twitter and Facebook, millions of user's daily share and exchange their opinions about different issues like products, events, persons or organizations on sites. Sentiment analysis on social users' data considered as a valuable analysis for automatically extract people opinions regarding some interested topic issues which enables to provide important information for informed decision making in different domains. With the noticed importance of sentiment analysis on social sites many applications and techniques are available.

Keyword: Features extraction, Context based sentiment analysis, Machine Learning

I Introduction

A. Data Mining

Data mining is also called as knowledge discovery in databases (KDD). It can be defined as the progression of estimating motivational, beneficial and unseen outlines from large bulks of data goods and finds the dealings among the outlines. Data mining job involves utilities for data of mathematics and Artificial Intelligence systems (AI). AI systems include neural networks and machine learning.

B. Sentiment Analysis

Sentiment analysis is the mission of categorizing positive or negative opinions, feelings in the text. SA is the multidisciplinary study which deeds methods from text mining, machine learning and natural language processing.

Sentiment analysis is the automated process of analyzing text data and classifying opinions as negative, positive or neutral.

Sentiment analysis different levels of scope

Document Level Sentiment analysis: In document level sentiment analysis it is required to extract informative text for inferring sentiment of the whole document. It is difficult for learning methods to deal with objective statements which can be rendered by subjective statements. This may lead to complicate further for document categorization task with conflicting sentiment. Sentiment of a complete document or paragraph.

Sentence Level Sentiment analysis: In this type polarity of the sentence can be given by three categories as positive, negative and neutral. It is a challenging area to deal with identification features indicating whether sentences are on-topic which is kind of co-reference problem. Sentiment analysis obtains the sentiment of a single sentence.

Feature Level Sentiment analysis: Feature level focuses on product features which are defined as product attributes or components. Feature based sentiment analysis means analysis of such features for identifying sentiment. In this approach positive or negative opinion is identified from the already extracted features.

C. Types of Sentiment Analysis

There are many types of sentiment analysis and SA tools systems that focus on polarity (*positive*, *negative*, *neutral*) to systems that detect feelings and emotions (*angry*, *happy*, *sad*, etc) or identify intentions (e.g. *interested or not interested*).

II Sentiment Analysis Process and Different methods

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. In the analysis process

- **Import Data:** We read in the comma separated file we downloaded from the Kaggle Datasets. We shuffle the data frame in case the classes are sorted.
- **Preprocessing:** The pre-processing steps help convert noise from high dimensional features to the low dimensional space to obtain as much accurate information as possible from the text. Preprocessing data can consist of many techniques depending on the data and the situation. Tokenization, Stemming, Stopwords, Negations, Lowercasing, Normalization, Lemmatization, Noise removal, part-of-speech tagging.

  Tokenization: Tokenization is the process of converting text into tokens before transforming it into vectors. It is also easier to filter out unnecessary tokens. In this case we are tokenizing the reviews into words.

  Stop words: Stop words are words which are filtered. Some examples of stop words are: “a”, “and”, “but”, “how”, “or”, and “what”. We removed these stop words [6].

  Negations: “This video game was not so good” this type of syntax will result in positive
  Outcome though the meaning denoted negative emotion. To overcome this issue, we have checked for this type of syntaxes (not good/ not so good, not bad/ not so bad) replaced them with their respective positive/ negative words [6].

  Lowercasing: Lowercasing all the text data, although commonly overlooked, is one of the simplest and most effective form of text preprocessing.

  Normalization: For example, the word “gooood” and “gud” can be transformed to “good”, its canonical form.

  Lemmatization: Lemmatization is related to stemming, differing in that lemmatization is able to capture canonical forms based on a word's lemma. For example “better” convert into “good”.

  Stemming: Stemming techniques put word variations like “great”, “greatly”, “greatest”, and “greater” all into one bucket, effectively decreasing entropy and increasing the relevance of the concept of “great”.

  Noise removal: This includes punctuation removal, special character removal, numbers removal, html formatting removal, domain specific keyword removal, source code removal, header removal and more.

  POS Tagging: Part-of-speech tagging aims to assign parts of speech to each word of a given text (such as nouns, verbs, adjectives, and others) based on its definition and its context.
- **Feature Selection & Feature Extraction:** Feature Selection in four main categories NLP or heuristic based, Statistical, Clustering based, Hybrid. Feature Extraction Techniques are Lexicon Based Sentiment Analysis, Aspect Based Sentiment Analysis, Corpus Based Sentiment Analysis, Semantic

Sentiment Analysis, Dictionary Based Sentiment Analysis and Context Based Sentiment Analysis. Sub task of feature extraction Bag of Words, TF-IDF Term Frequency-Inverse Document Frequency, word embedding, Doc2vec.

Lexicon Based Sentiment Analysis: Sentiment lexicon contains words with their sentiment orientation values. We construct the new sentiment lexicon for our approach based on three existing lexicons such as Sentistrength, SentiWordNet and Opinion lexicon [1].

Aspect Based Sentiment Analysis: In aspect category determination task, we have identified the aspect category of the aspect term extracted from given predefined set of aspect categories [2].

Semantic Sentiment Analysis: It would be classified in to two types contextual semantic and conceptual semantic. Contextual semantic deals with considering about neighboring word. Conceptual semantic depends on outside knowledge such as ontologies and semantic network [1].

Dictionary Based Sentiment Analysis: Dictionary-based sentiment analysis works by comparing the words in a text or corpus with pre-established dictionaries of words. These dictionaries could be based around positive/negative words.

Context Based Sentiment Analysis: Sentiment analysis move towards content, concept and context-based text analysis of natural language text. Context may be defined as any information that can be used to characterize the situation of an object or entity. In textual form, context considers the environment of the selected term in a sentence. Words that surround the selected term are important to explain the meaning of the sentence. Local context can capture features relations in a sentence, in sentiment analysis, local context is widely used in word sense disambiguation and valance shifters. Context-based SA is process of applying context to traditional SA aiming to improvise accuracy of results [7]. For example: The word "*unpredictable"* may invariably imply a negative sentiment when used in the context of a car's engine but is always positive when used in the context of a thriller movie's plot [3].

Word sense disambiguation**:** Word sense disambiguation, in natural language processing (NLP), may be defined as the ability to determine which meaning of word is activated by the use of word in a particular context.

Valance shifters: A negator flips the sign of a polarized word (e.g., "I do not like it."). see lexicon::hash_valence_shifters[y==1] for examples. An amplifier (intensifier) increases the impact of a polarized word (e.g., "I really like it."). see lexicon::hash_valence _shifters[y==2] for examples. A de-amplifier (downtoner) reduces the impact of a polarized word (e.g., "I hardly like it."). see lexicon::hash_valence_shifters[y==3] for examples. An adversative conjunction overrules the previous clause containing a polarized word (e.g., "I like it but it's not worth it."). see lexicon::hash_valence_shifters[y==4]

Bag of Words: The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

TF-IDF: TF-IDF is another way to convert textual data to numeric form, and is short for Term Frequency-Inverse Document Frequency. The vector value it yields is the product of these two terms; TF and IDF.

$$TF(t,d) = \frac{number\ of\ times\ term(t)\ appears\ in\ document(d)}{total\ number\ of\ terms\ in\ document(d)}$$

$$IDF(t,D) = \log\left(\frac{total\ number\ of\ documents(D)}{number\ of\ documents\ with\ the\ term(t)\ in\ it}\right)$$

Word embedding: When applying one-hot encoding to the words in the tweets, we end up with sparse vectors of high dimensionality (here the number of words).

Doc2vec: Doc2vec is to create a numeric representation of a document, regardless of its length. Doc2Vec's learning strategy exploits the idea that the prediction of neighboring words for a given word strongly relies on the document also.

- **Classification Algorithm:** This step expresses sentiment polarities as one of three options: positive, negative, or neutral [4]. Machine learning is further divided into two category namely supervised and unsupervised learning. Supervised classification algorithms are probabilistic classifier, linear classifier, decision tree and rule based classifier. Supervised learning technique is based on labeled dataset which is provided as input to train the model and this model is applied to test data to generate output. Machine learning-based classifiers including Calibrated Classifier (CC), Support Vector Classifier (SVC), AdaBoost (ADB), Decision Tree Classifier (DTC), Gaussian Naive Bayes (GNB), Extra Trees Classifier (ETC), Random Forest (RF), Logistic Regression (LR), Stochastic Gradient Descent Classifier (SGDC), and Gradient Boosting Machine (GBM), Support Vector Machines (SVM), and Naive Bayes (NB), K-Nearest Neighbors, Linear Regression, K-means Clustering are trained on dataset. Deep Learning Algorithms including Long short-term memory (LSTM), Backpropagation, Feedforward Neural Networks(FNN), Convolution Neural Networks(CNN), Recurrent Neural Networks(RNN), Recursive Neural Network, AutoEncoders, Deep Belief Networks and Restricted Boltzmann Machines.

- **Testing Accuracy:** When evaluating the sentiment of a given text document training a sentiment scoring system. Different performance metrics are available to measure the performance of classification systems like sensitivity, precision, F-measure, accuracy and specificity, recall. These performance metrics are generally used to analyses the performance of different models.

  Accuracy: Measures how many texts were predicted correctly (both as belonging to a category and not belonging to the category) out of all of the texts in the corpus.

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

  Sensitivity (Recall or True positive rate): Measures how many texts were predicted correctly as belonging to a given category out of all the texts that should have been predicted as belonging to the category. We also know that the more data we feed our classifiers with, the better recall will be. It is a ratio of true positive to the sum of true positive and false negative.

  Recall = TP / (TP+FN)

  Specificity (True negative rate): Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

  Specificity = TN / (TN+FP)

  Precision: measures how many texts were predicted correctly as belonging to a given category out of all of the texts that were predicted *(*correctly and incorrectly*)* as belonging to the category. It is a ratio of true positive to the sum of true positive and false positive. Test specificity (Precision) is the test's ability to correctly recognize those that do not have a disease (true negative rate).

Precision = TP / (TP+FP)

False Positive Rate: False positive rate is defined as the number of incorrect negative predictions divided by the total number of negatives.

FPR = FP / (TN+FP)

F-measure: The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is

$F = 1 / (\alpha\ 1/P + (1- \alpha)\ 1/R)$
where the weight $\alpha \in [0, 1]$.
In above equations

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

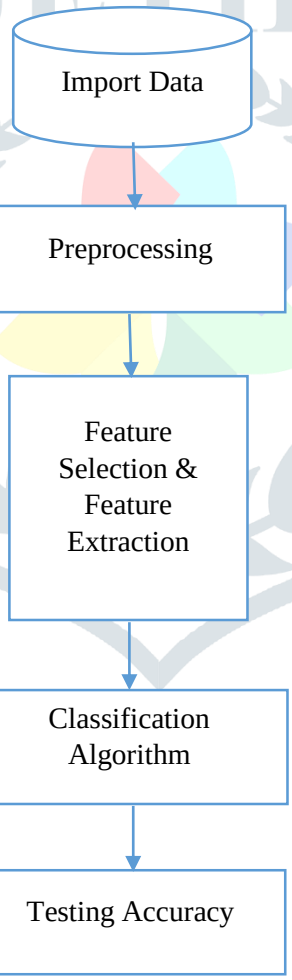

Fig 1: Sentiment Analysis Process

III Application of Data Mining

- The commercial product areas
- Brand monitoring
- Competitor monitoring
- The political areas
- The stock market and stock forecast

IV Advantages of Data Mining

- Adjust marketing strategy
- Measure ROI of your marketing campaign
- Develop product quality
- Improve customer service
- Crisis management
- Lead generation
- Sales revenue
- Predict accurate election result

Conclusion

In this paper we have studied the sentiment analysis using different methods based on the feedbacks of twitter datasets regarding. Context based sentiment analysis is process of applying context to traditional sentiment analysis aiming to improvise accuracy of results. Our results are a compelling evidence that the proposed model has high classification accuracy in predicting instances form the three classes (Positive, Negative, and Neutral).

Acknowledgement

I wish to warmly thank my guide, **Prof. Pragnesh Patel** for all her diligence, guidance, encouragement, inspiration and motivation throughout.

References

[1] [Vallikannu Ramanathan, T.Meyyappan]" Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism" [IEEE] (2019)

[2] [Ganpat Singh Chauhan and Yogesh Kumar Meena] "YouTube Video Ranking by Aspect-Based Sentiment Analysis on User Feedback" [Springer] (2019)

[3] [Srishti Sharma, Shampa Chakraverty, Akhil Sharma]" A Context Based Algorithm for Sentiment Analysis" [IEEE] (2018)

[4] [Shadi Abudalfa and Moataz Ahmed]" Open Domain Context-Based Targeted Sentiment Analysis System" [IEEE] (2018)

[5] [Oumayma El Ansari, Jihad Zahir, and Hajar Mousannif]" Context-Based Sentiment Analysis: A Survey" [Springer] (2018)

[6] [Shandro Chakraborty, Iftekharul Mobin, Abhijeet Roy, Mobtasim Hasan Khan]" Rating Generation of Video Games using Sentiment Analysis and Contextual Polarity from Microblog" [IEEE] (2018)

[7] [Nor Nadiah Yusof, Azlinah Mohamed, and Shuzlina Abdul-Rahman]" A Review of Contextual Information for Context-Based Approach in Sentiment Analysis" [International Journal of Machine Learning and Computing, Vol. 8, No. 4, August 2018] (2018)

[8] [Akshi Kumar, Geetanjali Garg]" Systematic literature review on context-based sentiment analysis in social multimedia" [Springer] (2019)

[9] [Sanjay Goswami, Satrajit Nandi and Sucheta Chatterjee]" Sentiment Analysis Based Potential Customer Base Identification in Social Media" [Springer] (2019)

[10] [TarekKanan, OdaiSadaqa, AmalAldajeh]" A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media" [IEEE] (2019)

[11] [Marius Ngaboyamahina, Sun Yi]" The Impact of Sentiment Analysis on Social Media to Assess Customer Satisfaction Case of Rwanda" [IEEE] (2019)

[12] [Xi Wang, IadhOunis, and Craig Macdonald]"Comparison of Sentiment Analysis and User Ratings in Venue Recommendation" [Springer] (2019)

[13] [Wei Zhang, Yue Zhang, and Kehua Yang]"Optimizing Word Embedding for Fine-Grained Sentiment Analysis" [Springer] (2019)

[14] [Zhang Feng]"Hot news mining and public opinion guidance analysis based on sentiment computing in network social media" [Springer] (2018)

[15] [Xin Xie, Songlin Ge, Fengping Hu, MingyeXie, Nan Jiang]"An improved algorithm for sentiment analysis based on maximum entropy" [Springer] (2017)

[16] [Heba Hakh, Ibrahim Aljarah, Bashar Al-Shboul]" Online Social Media-based Sentiment Analysis for US Airline companies" [The University of Jordan, Amman, Jordan. 25-27 April 2017] (2017)

[17] [Yun Xue, Xin Chen, Hongya Zhao, Xin Lu, Xiaohui Hu, Zhihao Ma]" A novel feature extraction methodology for sentiment analysis of product reviews" [Springer] (2018)

[18] [M.Trupthi, Suresh Pabboju, G.Narasimha]" Improved Feature Extraction and Classification - Sentiment Analysis" [IEEE] (2016)