REVIEW OF SENTIMENT ANALYSIS PAPERS

1. **Sentiment analysis on twitter data**

This paper gives the brief idea about two major tasks carried out while performing sentiment analysis of twitter data. First being POS specific prior polarity features. Secondly it describes about the use of a tree kernel to obviate the need for tedious feature engineering. Twitter is one of the microblogging website as it is used to post real time messages about different kinds of opinions on variety of topics.

In this paper, authors built models for classifying tweets into positive, negative and neutral sentiments. They considered 2 tasks: a binary task which segregates positive and negative classes and the other one that segregates into positive, negative and neutral classes. They experimented with three types of models: unigram, a feature and a tree kernel based model. The feature based model deals with the past features and also proposing new features. In tree kernel based model a new tree representation for tweets is designed.

The unigram model achieves over 20% accuracy for both the tasks. The feature model which used 100 features achieves similar accuracy as unigram that used 10,000 features. The tree model outperforms both the models by a significant margin. In this paper, the tweets are collected in streaming fashion and thus represents true sample of actual data. There are two resources available first being the hand annotated dictionary for emotions and other being the acronym dictionary that is collected from the web.

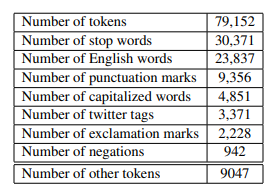


Fig 1: Statistics about the data

The above table gives the statistics about the data. In this data, approx 38.3% of the tokens are stop words, 30.1% are found in Wordnet, 1.2% are negation words.etc. For providing the polarity score, they have used Dicionary and extended it using WordNet. The dictionary assigns every word a score which ranges from 1(Negative) to 3(Positive). Then normalize it , following results can be derived:

|  |  |
| --- | --- |
| **Normalization value** | **Output** |
| X<0.5 | Negative |
| X>0.8 | Positive |
| otherwise | Neutral |

Then in the results they compared all the models and experiments performed. This can be explained as follows:

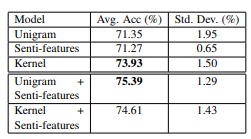
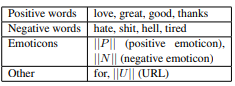
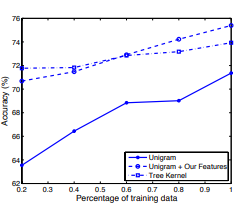
 

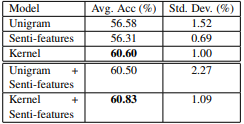
Fig 2: Avg and std dev for test accuracy for the 2 Fig 3: List of top unigram features for 2-way task

way classification task using different models

In unigram features they had used information gain as the attribute evaluation metric. Fig 2 gives the average and standard deviation for test accuracy for the 2-way classification task using different modules like unigram, tree kernel, senti- features etc. In fig 3, list of unigrams that appears as top 15 unigram features is highlighted. Positive and negative words are at the top.



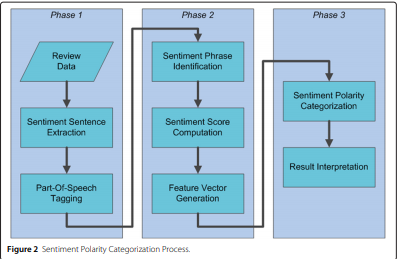
In the above snippet, they have compared the accuracy on the training data for unigram, tree kernel model. This is the learning curve for 2-way classification task. Then they also worked for 3-way classification task i.e combining two models and experimenting.



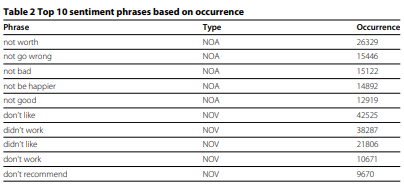
So it was concluded that tree kernel and feature based approach outperformed unigram. For the feature-based approach, they performed feature analysis which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags

1. **Sentiment analysis using product review data**

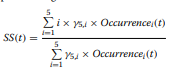
In this paper, the authors tried to tackle the problem of sentiment polarity categorization that is considered as one of the fundamental problem of sentiment analysis. The dataset is taken from amazon.com which consisted of data from February 2014 to April 2014. After that they performed the experiment for both sentence level and review level categorization. There are many flaws in the process of sentiment analysis. First flaw is that as people can freely post any content, the quality of their contents cannot be guaranteed. Examples spam. Secondly ground truth of online data is not always correct or available.



Above fig depicts the proposed process for Sentiment Polarity categorization. There are 3 phases. First phase consists of reviewing the data followed by sentiment sentence extraction. Finally, parts of speech tagging are done on the extracted sentence. In phase 2, initially an algorithm is proposed and implemented for negation phrases identification. Then a mathematical approach is proposed for sentiment score computations followed by feature generation method. Finally, in the last phase, Two sentiment polarity categorization experiments are performed based on sentence level and review level. At the end performance of the three models was compared based on the results.

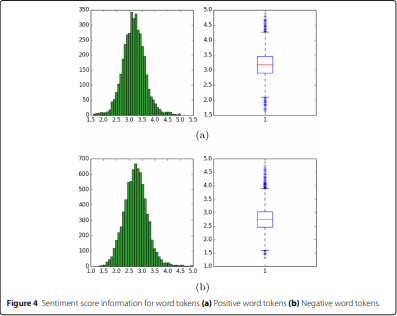
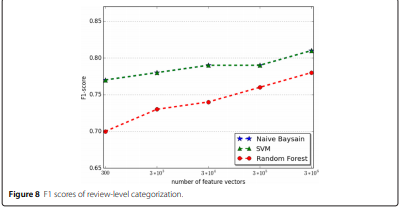


The above snippet gives the list of top 10 sentiment phrases based on occurrence. The Sentiment Score can be computed as follows:



Where occurrence(t) is t’s number of occurrence in i-star reviews i=1, 2, 3, 4, 5

Based on the above sentiment score following graph was observed:

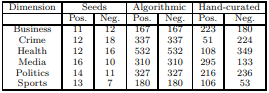
 

Experiments for both sentence-level categorization and review-level categorization have been performed. Software used for this study is scikit-learn, an open source machine learning software package in Python. The classification models selected for categorization are: Naïve Bayesian, Random Forest, and Support Vector Machine.

1. **Sentiment analysis on news and blogs**

The authors presented a system that assigns scores indicating positive or negative opinion to every entity in text corpus. The system consisted of sentiment identification phase and sentiment aggregation, scoring phase. Finally, the significance of scoring techniques over large corpus of news and blogs is evaluated. There are three steps carried out in the whole process:

* In the first step which is named as **algorithmic construction of sentiment dictionaries** they tracked the frequencies of adjectives with positive and negative connotations. They used sentiment alteration hop counts to get the polarity strength of the candidate terms eliminating the ambiguous terms.
* Second step is **Sentiment Index Formulation.** In this step they constructed statistical index that reflects the significance of sentiment term juxtaposition. They provided the technique of using juxtaposition of sentiment terms and entities.
* Final step is **Evaluation of significance.** They provided statistical evidence of their system by comparing the model with the real world classes like performance of stock market, seasonal effects etc. The result tells that there were positive correlations that proved that the sentiment analyzer accurately measured public sentiments.

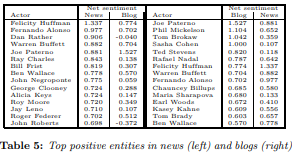
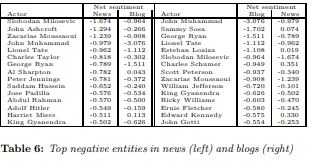


The table above shows the sentiment dictionary composition for adjectives. It tells the composition of the algorithmically generated and curated sentiment dictionaries for each class of adjectives.

The second step as discussed earlier computes the sentiment index. The ides followed is that they reversed the polarity of the sentiment word wherever it was preceded by a negation. Thus for every word the polarity was adjusted according to the modifier. Like for example, not good = -1, good = +1, very good = +2. These prevented the problem of duplicates in the text. There are basically two important terms: one is polarity and the other is subjectivity. Polarity is the sentiment associated with the entity either positive or negative. It gives the percentage of positive sentiment to the total number of sentiment references. Subjectivity tells the ratio of sentiment to frequency of occurrences. This is shown as below:

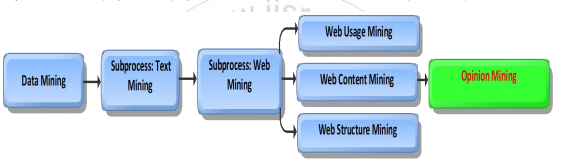
Then they compared the results obtained on newspapers and blogs. They generated top 15 positive and negative entities from each of news and blogs articles in the year 2006.

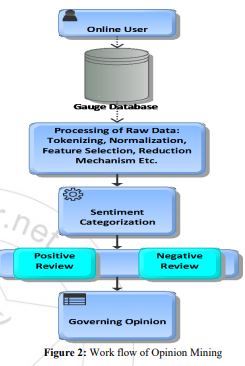
From these tables it can be interpreted that persons like Warren Buffet is present in both the list of news and blogs in positive entities. This is because he is a well-known American investor. Similarly, in the negative entities John A. Muhammed is present as he has many criminal records. But there are also few limitations as few persons who are referred as positive in news are referred as negative in blogs.

1. **Study on sentiment analysis: Methods and tools**

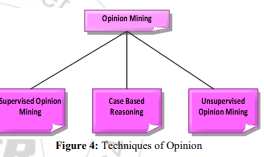
This paper focusses on different kinds of sentiment analysis methods and tools. Sentiment analysis is defined as a process of extraction of features by considering the notions of others regarding product, service, experience. The sentiment analysis tool applied on an item to perform series of operations on it. Sentiment analysis is also termed as Opinion mining as there are massive amount of opinions based on which decisions are made. The hierarchy of data mining is given below:



As we are moving towards new technologies, amount of data involved is also increasing day by day. To maintain the working of data smoothly there is definitely need of automation in every field. So automated analysis systems play the key role in analysis, summarization and classification of data to store huge amount of data. Text mining is the approach which uses different fields like machine learning, statistics, information retrieval. Next step is web mining. It is the subset of text mining which is used to mine unstructured web text to convert it into structured text. So the main aim of opinion mining is to make the automated machine to read and categorize emotions. The dataset contained various types of product reviews and movie reviews extracted from Flip-kart and IMDB webpage.



The above figure gives the workflow of opinion mining. Initially, the pre-processing is done which includes tokenization, normalization, feature selection. Tokenization is the process of splitting up text into tokens after removing white spaces, punctuations. Normalization is the process of converting whole document into one similar value either lowercase or uppercase. Then feature extraction deals with the extraction of features (Term Frequency, POS, Term co-occurrence, Opinion words, Syntactic dependency). Second step is sentiment categorization. This includes Opinion Retrieval, classification and summarization. Then next step is review classification. It segregates as positive and negative review. The techniques used for opinion mining are supervised mining, case based reasoning and unsupervised opinion mining.



The supervised opinion mining is used when the target value is known. It gives relationship between independent and dependent attributes. Propagation algorithm is used to extract Product features and sentenced words. In unsupervised learning, there is no definite targeted output. Clustering is the method used for this purpose. Objects in one cluster is different from the objects in other clusters. Case Based Reasoning is tool of computer reasoning and helps in cracking the problems in closest way to real time scenario. The solutions of all cases are stored in CBR warehouse known as knowledge base.

1. **Deep convolutional neural network for sentiment analysis of short texts**

In this paper, the authors proposed new deep convolutional neural network that performed the sentiment analysis of short texts. They applied this approach for two corpora of two different domains: Stanford Sentiment Tree-bank which contains movie reviews and the other one Stanford Twitter Sentiment corpus which has Twitter messages. For the earlier one accuracy achieved was 85.7% and for the later one it was 86.4%. The proposed model was named as Character to sentence Convolutional Neural Network (CharSCNN). It uses two convolutional layers to get the relevant features from sentences.

The network takes sequence of words as input and then passes it through the sequence of layers to extract the features from character level to sentence level. The architecture contains 2 convolutional layers that allows it to handle words or sentences of any size.

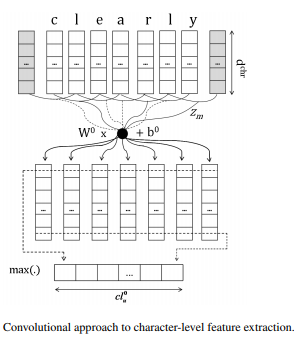
The first layer of the network is used to transform words into real valued feature vectors which is also known as embeddings. This is composed of two sub vectors: word-level embedding and the other is character level embedding. Word level embeddings are used to capture syntactic information and character level embeddings are used to capture morphological and shape information. In Word level embedding each column corresponds to the i-th word in the vocabulary. Word w can be transformed into its corresponding word-level embedding by:



Where vw is vector of size | V wrd | that has value 1 at index w and otherwise 0. To extract morphological and shape information, Character-level embedding is used. For this, word of M characters is converted into a character emebedding. This is represented as follows:



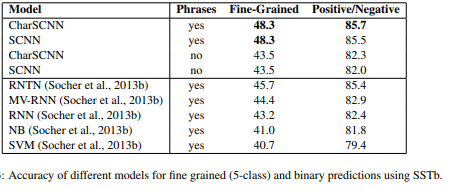
Where vc is vector of size | V chr | that has value 1 at index w and otherwise 0



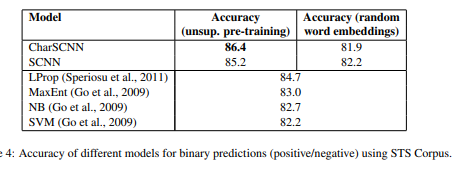
The convolutional layer computes the j-th element of the vector that is nothing but character-level embedding of w,



Where W0 is weight matrix of the convolutional layer. B0 is the bias value or learning parameter. After all this processing’s the results obtained were:



These results suggest that the character-level information is not much helpful for sentiment prediction in the SSTb corpus. Using phrases as training examples allows the classifier to learn more complex phenomena, since sentiment labeled phrases give the information of how words (phrases) combine to form the sentiment of phrases (sentences).



Here character level information over performed for twitter data. It had more accuracy as compared to others. Initializing word-embeddings using unsupervised pre-training gives an absolute accuracy increase of around 4.5 when compared to randomly initializing the word-embeddings.