Paper 13 | Approaches for Sentiment Analysis on Twitter: A State-of-Art study by Harsh Thakkar et. al – published in 2015

Abstract

1. This paper discusses the state-of-art sentiment classifiers that were applied to analyse Twitter, which involves surveying various lexical, machine learning and hybrid approaches

Background

1. The research on sentiment analysis has mainly focused on two things:
   - Subjective vs Objective
   - Polarity (positive vs negative) of subjective texts
2. Sentiment extraction is employed on Twitter posts using the following techniques:
   - Lexical analysis
   - Machine learning based analysis
   - Hybrid analysis
3. Lexical analysis
   - Governed by the use of a dictionary consisting pre-tagged lexicons
   - The input text is converted to tokens by Tokeniser
   - Every new token is match for the lexicon in the dictionary
   - A positive sentiment match, then score is added to the total score for the input text and vice versa, a negative match will lead to decrease in the total score
   - The classification of a text depends on the total score it achieves
   - The accuracy of this approach seems to rely heavily on the different lexicons dictionary being used. Also, the performance of this approach degrades drastically with the exponential growth of the size of dictionary

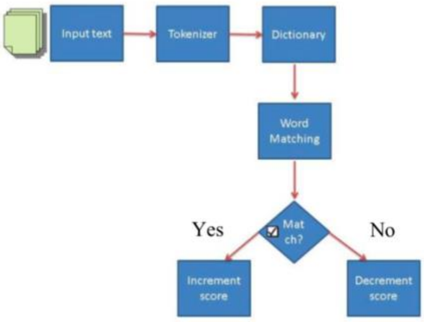

Fig. 1. Working of a lexical technique.

4. Machine learning analysis
   - Mostly supervised learning techniques
   - Five stages:
     1. Data collection
     2. Pre-processing (cleaning) data
     3. Finalise training data (adding special features)
     4. Training classifiers (classification)
     5. Accuracy, performance tuning and predictions

- In this approach, the key to accuracy is the selection of appropriate features. This approach faces the challenges of designing a classifier, availability of training data, and the correct interpretation of an unforeseen phrase

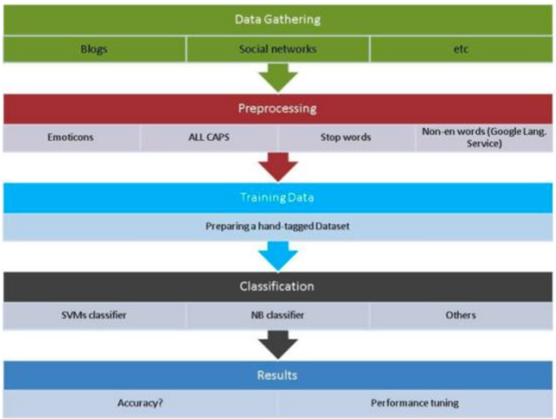

Fig. 2. Steps involved in the machine learning approach.

5. Hybrid analysis
    - Combining the accuracy of a machine learning approach with the speed of lexical approach

Summary

1. Comparison of all approaches has showed that best results have been observed from machine learning approaches and least by lexical approaches. However, without any proper training of a classifier, the machine learning approach results may deteriorate drastically

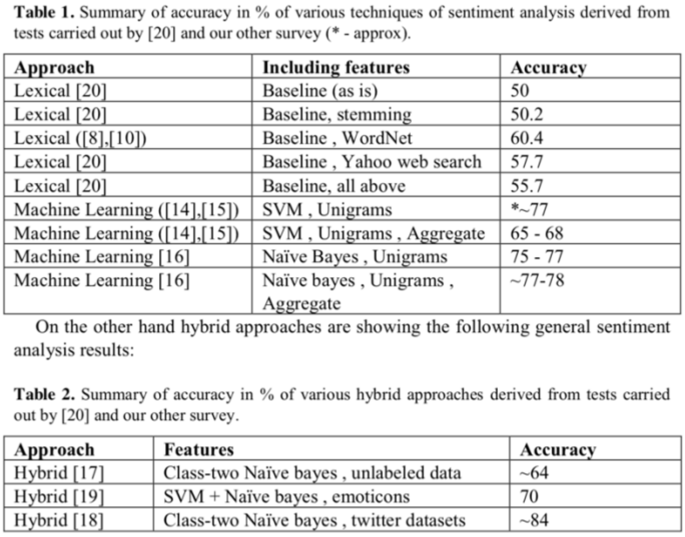

Table 1. Summary of accuracy in % of various techniques of sentiment analysis derived from tests carried out by [20] and our other survey (* - approx).

Approach	Including features	Accuracy
Lexical [20]	Baseline (as is)	50
Lexical [20]	Baseline, stemming	50.2
Lexical ([8],[10])	Baseline , WordNet	60.4
Lexical [20]	Baseline , Yahoo web search	57.7
Lexical [20]	Baseline, all above	55.7
Machine Learning ([14],[15])	SVM , Unigrams	*~77
Machine Learning ([14],[15])	SVM , Unigrams , Aggregate	65 - 68
Machine Learning [16]	Naïve Bayes , Unigrams	75 - 77
Machine Learning [16]	Naïve bayes , Unigrams , Aggregate	~77-78

On the other hand hybrid approaches are showing the following general sentiment analysis results:

Table 2. Summary of accuracy in % of various hybrid approaches derived from tests carried out by [20] and our other survey.

Approach	Features	Accuracy
Hybrid [17]	Class-two Naïve bayes , unlabeled data	~64
Hybrid [19]	SVM + Naïve bayes , emoticons	70
Hybrid [18]	Class-two Naïve bayes , twitter datasets	~84

Conclusion

1. Machine learning approaches have been good in delivering accurate sentiment results
2. Lexical approach is a ready-to-go and doesn't require any prior information or training whereas machine learning requires a well-designed classifier, huge amount of training data sets and performance tuning prior to deployment
3. Hybrid approach has shown great performance too although their performance seems to be worse on trigrams