A Comparative Analysis of Sentiment Analysis Tools for Data Mining Systems

Emer Coughlan
Department of Computer & Software Eng.
Faculty of Engineering & Informatics
Athlone Institute of Technology - Ireland
e.coughlan@research.ait.ie

David Scott
Department of Computer & Software Eng.
Faculty of Engineering & Informatics
Athlone Institute of Technology - Ireland
davidscott@ait.ie

Mark Daly
Department of Computer & Software Eng.
Faculty of Engineering & Informatics
Athlone Institute of Technology - Ireland
mdaly@ait.ie

Introduction

Over the past few years the rise in big data is something that has gained a lot of traction and attention, the sheer volume of data produced by humans is increasing rapidly each year.

The growth in social media over the past few years has changed how we interact and how we connect globally, as such there is a rising need to ascertain user sentiment. Sentiment Analysis (SA) also known as Opinion Mining (OM) is a means of understanding the emotion, opinion or sentiment of a piece of text and is an ongoing field of research. Natural Language Processing (NLP) is used to assign or identify meaning to written or spoken language or use in further analysis with machine learning.

Keywords: Sentiment Analysis, Natural language processing, social media, Classification, Big Data

Experimental Setup

Data Set: The data set used was a mix of headlines and by-lines from a selection of the most popular newspapers around the world as well as select twitter data based on choice input criteria

Pre-processing: To clean the data a few processes such as tokenization, stop word removal, regex pattern matching to remove certain special characters such as http, @. A negations dictionary to convert contractions back to their original word meanings like couldn't too could not. The data was all changed from uppercase to lower case and numbers were removed.

Annotation : As the data stands it can be difficult for the classifiers to decide what the polarity of the data might be, whether it be negative, positive or neutral. In order to annotate the polarity there was two different open source tools used, VADER (Valence Aware Dictionary and sentiment Reasoner) which is a python library used to find the sentiment expressed in social media primarily, Stanford NLP which is a Java library

Classification : When deciding to choose the best classifiers for sentiment analysis in supervised learning it was important to look at what classifiers were most suited to the type of data been processed. From previous work Logistic Regression, Random Forest and Linear SVC seemed to be among the most widely used classifiers

Results

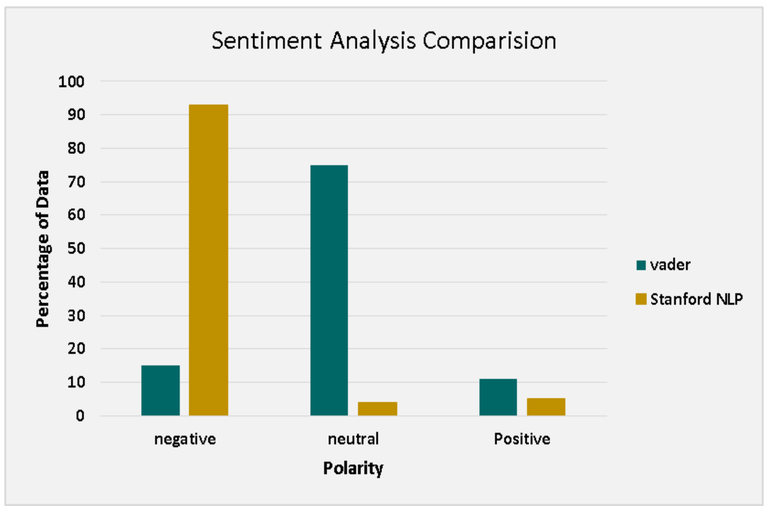

Fig 1: Comparison of polarity classification from VADER and Stanford

The data collected and processed was ran through two open source tools mentioned in the experimental set up. As seen in Figure 1 VADER classified the majority of data as neutral in polarity while Stanford classified it as negative. VADER is more generic than Stanford for categorizing sentiment.

	Precision	Recall	F1-Score
Neg	0.98	.98	0.98
Neu	0.81	0.84	0.82
Pos	0.61	0.70	0.65
Avg/Total	0.97	0.96	0.96

Fig 2: Classification report of Linear SVC with Stanford NLP

	Precision	Recall	F1-Score
Neg	0.85	0.84	0.85
Neu	0.96	0.96	0.95
Pos	0.88	0.77	0.82
Avg/Total	0.92	0.92	0.92

Fig 3: Classification report of Linear SVC with VADER

In Fig 2-3 it can be seen that the metrics of precision recall and F1 score all have a rather high percentage for Negative and Neutral in their respective majorities. The reason for this is most likely due to the fact that data is unbalanced and there is not enough test/train data for the lower percentages in each model as seen in Fig 1

Conclusion

The results above showed the difference in how the two analysers classified the different data both with a skewedness towards one category in the three point Likert scale. The Stanford data produced a higher accuracy than that of the Vader.

Due to the negative nature of data one would have to delve deeper into the data and start looking at the level of negativity perhaps in a five or seven point Likert Scale. This would provide the intensity of the negativity within a tweet or headline