Opinion mining from Twitter messages

Nada Radwan Mohsen Supervisor: Professor Diana Inkpen
School of Electrical Engineering and Computer Science
University of Ottawa, Canada

Abstract

Automatic extraction of information related to particular topics from social media websites can be facilitated by programs that act like search engines. Extending these programs to show statistics of the percent of people for and against a particular idea or product is very useful particularly in the areas of marketing and decision making. This research project aimed to identify public opinion as well as obtain statistics from social media participants (particularly Twitter messages) regarding their opinions on a variety of topics. This was done through the retrieval of a collection of Twitter messages about certain topics, and then the use of machine learning involving the program Weka to classify the messages into positive, negative, objective, or neutral. Weka was trained to determine the polarity of messages through the use of pre-labeled training data. Advancements to the field of sentiment analysis are of great use and importance for companies who want to know about whether a tried software package is worth the investment or even individuals who are worried about buying one phone or the other. The world of opinion mining has been made a click away.

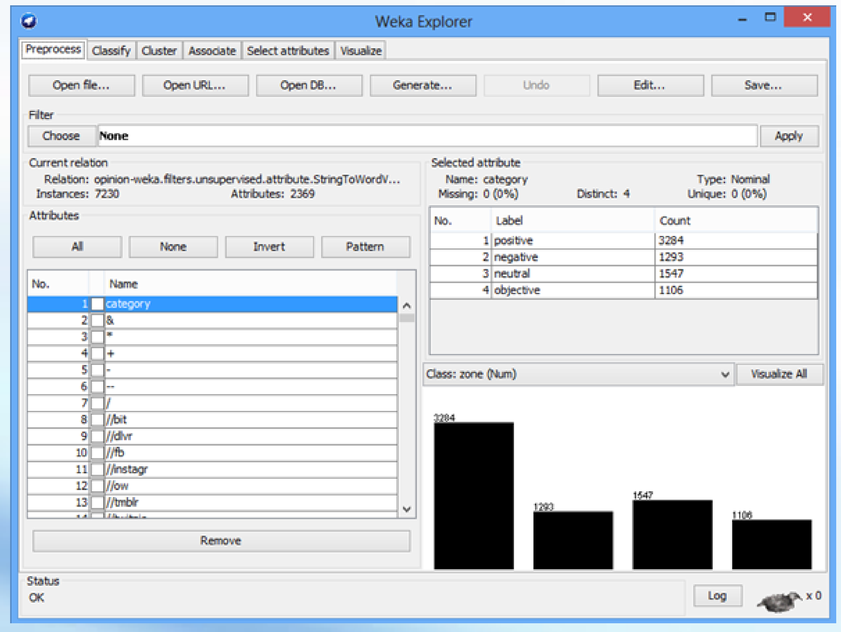
Introduction

The public's opinion dictates our understanding and perception of everyday subjects. We are constantly striving to learn about and discover different points of view regarding the best phone application, the latest car, or even the most supported political party. Companies would like to know what people think of their goods in order to maintain standards or improve the quality of their products and services. Because people are likely to reveal their opinion on social media websites, it would be great if this data could be automatically classified to serve readily as tool for decision making and marketing. The automatic technique used for such purposes is machine learning. Machine learning is a branch extending from computer science and statistics. It is the ability of a computer to program itself from experience together with the use of a variety of algorithms. These algorithms make use of statistical formulas to store and merge data. This research paper made use of the program Weka, which is a collection of machine learning algorithms. The program was trained to determine the polarity of twitter messages through the use of pre-labeled training data. Three learning classifiers were used and the percent accuracy of each assessed.

Methodology

We used supervised learning classifiers from Weka: Zero R, Naïve Bayes, and SVM (Support Vector Machine). We used 7230 annotated Twitter messages on a variety of topics from SemEval. Cross validation was used so that 90% of the data was used for training the classifier and 10% was used for testing. Training and testing data alternated to ensure that each message has been used for both purposes.

Weka classified the training data into the 4 classes (positive, negative, objective, and neutral).



Results

The classifiers used the training data to learn to identify the polarity of messages using different techniques. The results of the number of positive, negative, objective, and neutral messages produced by each classifier was averaged and the percentage accuracy with respect to the previously annotated polarity from SemEval was calculated.

Conclusion

The Naïve Bayes algorithm produced the most accurate classification (56.34% correctness) in an almost negligible amount of time. Even though SVM was close behind with 52.57% correctness, it took SVM 268.58 seconds to produce the results. Zero-R was quick but the results were inaccurate at 45.4% correctness. The percent accuracy of Zero-R is expected to be low as the nature of the technique involves grouping all the messages into the positive class. Therefore, it seems that for the case of Twitter message classification, the most efficient technique accuracy and time-wise was the Naïve Bayes.

Zero-R & Confusion Matrix

	F-Measure	Class
	0.499	Positive
	0.499	Negative
	0.499	Neutral
	0.499	Objective
Weighted Average	**0.454**	

Classified as →	a	b	c	d
a=positive	3284	0	0	0
b=negative	1293	0	0	0
c=neutral	1547	0	0	0
d=objective	1106	0	0	0

Naïve Bayes & Confusion Matrix

	F-Measure	Class
	0.702	Positive
	0.501	Negative
	0.308	Neutral
	0.49	Objective
Weighted Average	**0.563**	

Classified as →	a	b	c	d
a=positive	2484	285	276	239
b=negative	414	626	149	104
c=neutral	611	235	394	307
d=objective	283	60	193	570

SVM & Confusion Matrix

	F-Measure	Class
	0.67	Positive
	0.425	Negative
	0.331	Neutral
	0.408	Objective
Weighted Average	**0.526**	

Classified as →	a	b	c	d
a=positive	2410	286	392	196
b=negative	496	502	210	85
c=neutral	645	203	476	223
d=objective	364	81	248	413

References and Acknowledgement

Semeval 2012 Twitter Analysis Task ,
http://www.cs.york.ac.uk/semeval-2013/task2/

A Special Thanks to Professor Diana Inkpen for making this research possible.