# Towards SENTIEXTRACT: A Combination of OCR and Sentiment Analysis

Sumitra Pundlik
Associate Professor,
Computer Department
MIT College of Engineering
University of Pune, India

Varun Rambal
Student, Computer Engineering
MIT College of Engineering
University of Pune, India

Sonu Tayade
Student, Computer Engineering
MIT College of Engineering
University of Pune, India

Sonali Ramteke
Student, Computer Engineering
MIT College of Engineering
University of Pune, India

Pratigya Suri
Student, Computer Engineering
MIT College of Engineering
University of Pune, India

## ABSTRACT

Do you have a lot of unstructured data in image files? Are you interested in finding out the sentiment of those files? If you are SENTIEXTRACT is the perfect tool for you. In this paper, we have given an insight of our system (SENTIEXTRACT). Our system works on algorithms such as tesseract-ocr to convert image files to text files and naïve bayes classifier to find out the sentiments of these files.

In this system we are using a data set of movie reviews collected from IMDB. Giving this dataset as training dataset to our naïve bayes classifier, we have tried to achieve high accuracy for our system. Also, experimental results of how our system responds are shown for image files based different size and number of words.

## General Terms

Sentiment Analysis, Optical Character Recognition, Internet Movie Database

## Keywords

SENTIEXTRACT

## 1. INTRODUCTION

Prior to digitalization, companies recorded their data on paper. A non-ecofriendly and time consuming process. Imagine to try and find any useful information from rooms full of papers that have important information on them, digitalization has made this much easier. In this modern era companies have realized the importance of the data that they have accumulated throughout these years and are trying to make sense out of it. Company's track record, customer's feedback, companies share market reviews etc, can all be analyzed and the analysis can be used to improve the company's performance.

Opinion extraction or semantic classification is a related problem of studying the semantic orientation or polarity of words, which is also commonly known as Sentiment Analysis[1]. Its many forms are textual, audio, video etc.

In this paper we are talking about our system SENTIEXTRACT which can be directly used for finding out sentiments of data that are stored on paper. The prerequisite of this system is that the data on these papers must be converted to digital image files. Only image files can be given as input to our system. Optical Character Recognition[2] is the tool that we have used in our system to convert the digitally scanned image files to text file for the process of sentiment extraction, which would be carried out using the naïve bayed classifier[3].

This paper is organized as follows. In section 2, we have done the literature survey on Sentiment Analysis (SA) and Optical Character Recognition (OCR). In section 3, we concentrate on the working of SENTIEXTRACT and the dataset that we have used for training our naïve bayes classifier. Analysis of results shown by our system are compared in section 4 and the conclusion and future work is covered in section 5.

## 2. LITERATURE SURVEY

### 2.1 Sentiment Analysis

Various researchers are working on different problems of sentiment analysis.Such as, Jeonghee Yi et.al.[4] have worked on extraction sentiments from a given topic using natural processing techniques. Alexander Pak and Patrick Paroubek[5] have focused on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. Using this corpus, they have built a sentiment classifier, which is able to determine positive, negative and neutral sentiments for a document. Theresa Wilson et.al.[6] have worked on an approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. Many researchers have used information available on twitter as their dataset to create different systems such as, relationship between NFL betting and public opinions in blogs and twitter was proposed by Hong and Skein [7]. Twitter sentiments was linked with public opinion polls by O'Conner et.al [8]. Twitter sentiments were applied to predict election results by Tumasjan et.al [9]. Still being in its early days, Sentiment Analysis has many areas in which researchers have not yet looked upon.

### 2.2 Optical Character Recognition

Being a widely researched field, many researchers have worked on Optical Character Recognition(OCR). A complete OCR system for picture embedded text document for handheld devices was designed by Mollah Ayatullah Faruk et.al. [10]. A new method of OCR which is based on the pattern character recognition algorithms and uses hierarchical optimisation was proposed by Safronov K. et.al. [11]. A novel

Web image processing algorithm that aims to locate text areas and prepare them for OCR procedure with best results was talked about by Perantonis S. J.et. al. [12]. Character recognition process from printed documents containing Hindi and Telugu text was researched by Jawahar C. V. et.al. [13]. An OCR method that recognizes and extracts Arabic text from image files and converts it into format that can be edited was discussed by Mesleh Abdelwadood et.al. [14]. A comprehensive overview of the Tesseract OCR engine in which emphasis on aspects that are novel or at least unusual in an OCR engine was talked about by Smith Ray [15]. He has also proposed a new, accurate and robust skew detection algorithm based on a method for finding rows of text in page images [16].

# 3. SENTIEXTRACT

Our system boasts of two highly efficient algorithms namely, tesseract ocr algorithm for image to text conversion and naïve bayes classifier for sentiment extraction of the converted text. A single or multiple image files can be given as input to our system for calculation of their sentiment polarity.

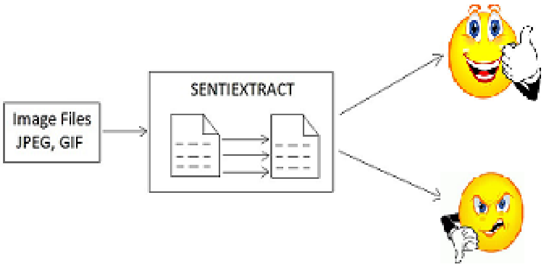

**Figure 1. Block Diagram of SENTIEXTRACT**

## 3.1 Extraction of Text From Picture Files

The image file is firstly converted to a text file using the tesseract ocr algorithm, in our system we are using the pyocr wrapper for this purpose. This wrapper invokes the tesseract algorithm and gives us the required text file, which is then used for further analysis. Image files of any size can be given as input to the system, but care should be taken as blurred or skewed images may not fetch you optimal results.

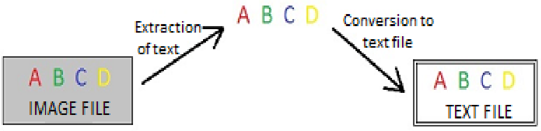

**Figure 2. Block Diagram of OCR**

Number of words a file many contain has no constraints, but the words should be of English language as it is the only language currently supported by our system.

## 3.2 Sentiment Extraction From Text

This text file is used as an input to the naïve bayes classifier which gives us the ratio of this document in two classes i.e., positive and negative. But before this text file can be given as an input to our classifier, our classifier needs to be trained. We have trained our classifier using the movie_reviews database given in the nltk corpora.

The documents passed to the naïve bayes classifier are represented as a bag of words (a bag is like a set that allows repeating elements). This is an extremely simple representation, it only knows which words are included in the document and how many times each word occurs, and throws away the word order.

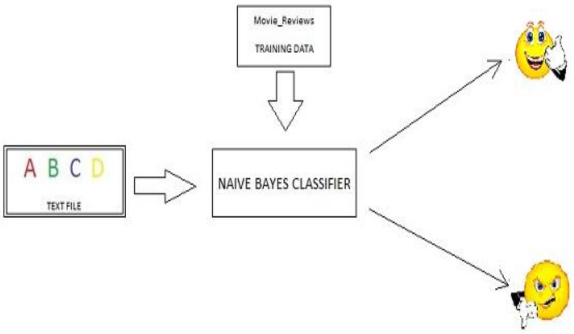

**Figure 3. Block Diagram of Sentiment Analysis**

After the classifier is trained it is ready to process the converted text file. For each file the classifier is first trained and then it classifies the files to its specific polarity. This help to increase the accuracy of the classifier but takes more time to execute.

## 3.3 Dataset

Our dataset represents an enhancement of the review corpus v1.0, it contains more reviews, and labels were created with an improved rating-extraction system. This data was first used in Bo Pang and Lillian Lee[17].

Within the folder "txt_sentoken" are the 2000 processed down cased text files used in Pang/Lee ACL 2004. The names of the two subdirectories in that folder are "pos" and "neg". which indicate the true sentiment of the component files.

Each line in each text file corresponds to a single sentence. Preliminary steps were taken to remove rating information from the text files, but only the rating information upon which the rating decision was based is guaranteed to have been removed. Thus, if the original review contains several instances of rating information, potentially given in different forms, those not recognized as valid ratings remain part of the review text.polarity_html.zip is the original source files from which the processed, labeled, and (randomly) selected data in review_polarity.tar.gz was derived.

This data consists of unprocessed, unlabeled html files from the IMDb archive of the rec.arts.movies.reviews newsgroup, http://reviews.imdb.com/Reviews. The files in review_polarity.tar.gz represent a processed subset of these files.

# 4. RESULTS AND ANALYSIS

The system configuration that we using to implement SENTIEXTRACT is memory :1 GB, number of processors : 1, Operating System : Linux 10.04

After performing several tests on our system we have derived the following results based on parameters such as size of input file, number of words in input file, time taken to execute the given input file.

From the above table we are able to analyze that as the file size increases for the same number of words the time taken by our system to execute also increases.

**Table 1. Analysis of Result 1**

| File size | 50 words (time taken) | 100 words (time taken) | 200 words (time taken) |
|---|---|---|---|
| < 1MB | 12 sec (85 KB) | 12 sec (157KB) | 12 sec (41 KB) |
| 1MB – <5MB | 19 sec (1.2MB) | 21 sec (1.5MB) | 21 sec (1MB) |
| >5MB | 2 min 10 sec (5MB) | 1 min 31 sec (5.1MB) | 2 min 30 sec (~5MB) |

This can also be represented graphically as shown below

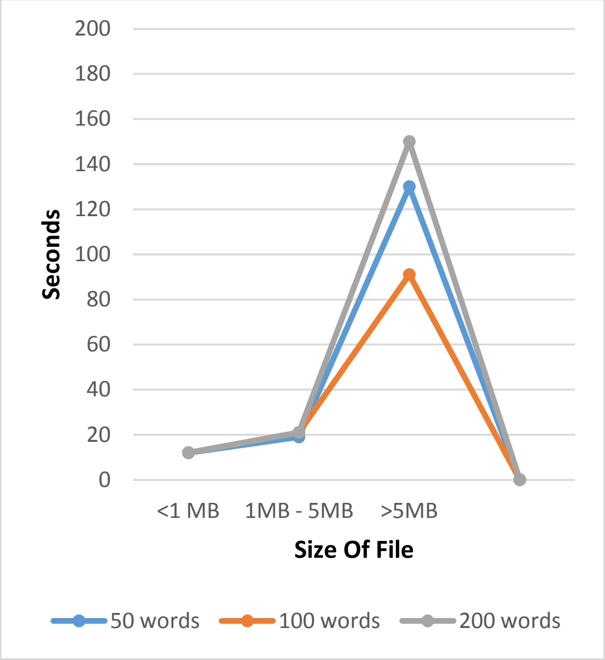

**Figure 4. Graphical Representation of Result 1**

Another analysis is shown below for multiple input files and the time required by our system to execute them. The analysis of the below table shows us that as the number of input files increases the time taken to execute them also increases.

Further on, this result has also been shown graphically.

**Table 2. Analysis of Result 2**

| No. of image files | Time Taken |
|---|---|
| 5 | 18 sec |
| 10 | 27 sec |
| 15 | 36 sec |
| 20 | 44 sec |
| 25 | 51 sec |
| 30 | 1 min 3 sec |
| 35 | 1 min 11 sec |
| 40 | 1 min 17 sec |
| 45 | 1 min 25 sec |
| 50 | 1 min 32 sec |

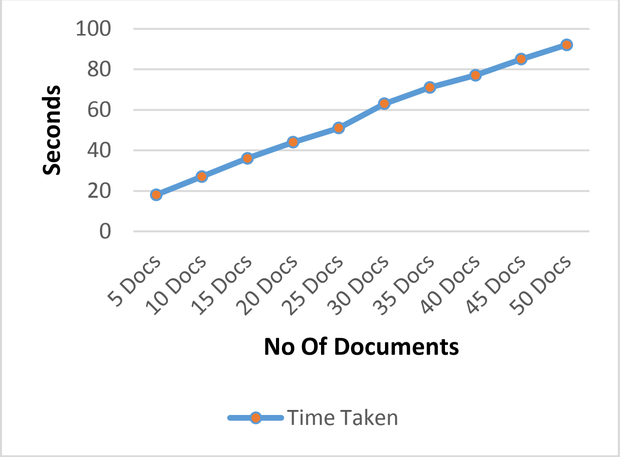

**Figure 5. Graphical Representation of Result 2**

## 5. CONCLUSION AND FUTURE WORK

In this paper we have gazed upon the work being done by different researchers in the field of Sentiment Analysis and Optical Character Recognition. Also we have given an insight of our system SENTIEXTRACT, we have described its working and have given the analysis of the results shown by our system.

As our system is still in its adolescence a lot of future work is required, better image recognition techniques as well as more efficient sentiment extraction methods can be used in our system.

## 6. REFERENCES

[1] Bing Liu,2012, Sentiment Analysis and Opinion Mining.

[2] Ravina Mithe, Supriya Indalkar and Nilam Divekar,2013, Optical Character Recognition, IJRTE, Volume-2, Issue-1, March 2013.

[3] A McCallum, K Nigam,1998, A comparison of event models for naive bayes text classification.

[4] Jeonghee Yi, 2003 , Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques.

[5] Alexander Pak, Patrick Paroubek,2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

[6] Theresa Wilson, Janyce Wiebe, Paul Hoffmann,2005, Recognizing contextual polarity in phrase-level sentiment analysis.

[7] Hong, Yancheng and Steven Skiena,2010, The Wisdom of Bookies? Sentiment Analysis vs. the NFL Point Spread,Proceedings of the International Conference on Weblogs and SocialMedia.

[8] O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith, 2010,From Tweets to Polls: Linking Text Sentimentto Public Opinion Time Series,Proceedings of the International AAAI Conference on Weblogs and Social Media.

[9] Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe,2010, Predicting elections with twitter: What 140 characters reveal about political sentiment, Proceedings of the International Conference on Weblogs and Social Media.

[10] Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri,2011, Design of an Optical Character Recognition System for Camera-based Handheld Devices, IJCSI, Vol. 8, Issue 4, No.1,July,2011.

[11] Kirill Safronov, Dr.Ing. Igor Tchouchenkov and Prof. Dr.Ing. Heinz Wörn,2007, Optical Character Recognition Using Optimisation Algorithms, Proceedings of the 9thInternational Workshop on Computer Science and Information Technologies, Ufa, Russia, 2007.

[12] S. J. Perantonis, B. Gatos and V. Maragos,2003, A novel Web image processing algorithm for text area identification that helps commercial OCR engines to improve their Web image recognition efficiency.

[13] C. V. Jawahar, M. N. S. S. K. Pavan Kumar and S. S. Ravi Kiran,2003, A Bilingual OCR for Hindi-Telugu Documents and its Applications.

[14] Abdelwadood Mesleh, Ahmed Sharadqh, Jamil Al-Azzeh, MazenAbu-Zaher, Nawal Al-Zabin, Tasneem Jaber, Aroob Odeh and Myssa'a Hasn,2012, An Optical Character Recognition, Contemporary Engineering Sciences, Vol. 5, No. 11,2012.

[15] Ray Smith, 2007,An Overview of the Tesseract OCR Engine, IEEE, 2007.

[16] R. Smith, A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation,Proc. of the 3rd Int. Conf. on Document Analysis and Recognition,IEEE, Vol. 21995.

[17] Bo Pang and Lillian Lee,2004, A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.