WWJMRD 2017; 3(9): 48-51
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

K.Deepa
Assistant Professor (Research Scholar), Department of CS, Nehru Memorial College, Puthanampatti, Trichy (DT), India

Dr.E.Kirubakaran
Research Supervisor, Department of Computer Science, Bharathidasan University, Thiruchirapalli,, India

Correspondence:
K.Deepa
Assistant Professor (Research Scholar), Department of CS, Nehru Memorial College, Puthanampatti, Trichy (DT), India

A Review on Sentiment Analysis Techniques

K.Deepa, Dr.E.Kirubakaran

Abstract
In this paper, the overviews of the Sentiment Analysis techniques are surveyed. It is challenging to understand the latest trends and summaries the state or general opinions about products due to the big diversity and size of social media data and this creates the need of automated and real time opinion extraction and mining. Mining online opinion is a form of sentiment analysis that is treated as a difficult text classification task. In this paper, we explore the role of text pre-processing in sentiment analysis, and report on experimental results that demonstrate that with appropriate feature selection and representation, sentiment analysis accuracies using support vector machines (SVM) in this area may be significantly improved. The level of accuracy achieved is shown to be comparable to the ones achieved in topic categorization although sentiment analysis is considered to be a much harder problem in the literature.

Keywords: Sentiment Analysis, Preprocessing, SA techniques

Introduction

Sentiment analysis in reviews is the process of exploring product reviews on the internet to determine the overall opinion or feeling about a product. Reviews represent the so called user-generated content, and this is of growing attention and a rich resource for marketing teams, sociologists and psychologists and others who might be concerned with opinions, views, public mood and general or personal attitudes.

With the increasing importance of social media information in every domain of today's digital age from algorithmic trading and product recommendations to politics, there is a tremendous amount of research work going on in the field of sentiment analysis and opinion mining which is taking us leaps and bounds with the advent of Big Data platforms and tools. The amount of data that could be gathered processed and stored cheaply and effectively is increasing at an exponential rate with the advent of Hadoop and other related massively parallel platforms and tools. Our work aims at studying the importance of pre-processing in the age of big data where storage and processing of unstructured data is as simple as processing structured data.

So why do we have to pre-process the data if Hadoop and other big data tools support handling unstructured data effectively? If required what kind of pre-processing are we talking about and how different it is from the pre-processing that we do in a regular KDD process? What kind of tools work well in such a scenario and how it is done effectively on such a huge volume of data? Since most of the tools in the Hadoop Ecosystem fundamentally works on the basis of the Map Reduce paradigm (which is a batch processing model), how well do they handle the pre-processing of data that is arriving at a faster rate, like Twitter feeds or posts from Facebook users? How do we handle the velocity part of the Big Data problem? Are the tools of the past no longer valid for these purposes due to the huge volume, variety and velocity of data? These are some of the questions that we are trying to answer.

The objective of this work is to identify the best framework or set of tools to pre-process the data from a social networking site like twitter. Even though most of the algorithms for mining big data are found to be supporting unstructured data and also found to be robust to variations in data formats and structure, we stress the importance of pre-processing data as its advantages are found to be many-folded. First it lets us understand the unstructured data that we are dealing with in a better way. Second it helps in dimensionality reduction thereby eliminating a lot of unnecessary features from being handled and in some cases making In-Memory processing of data possible resulting in a huge reduction of I/O overhead. Third it allows answering the Value and Veracity part of the Big Data paradigm. Finally it allows us to fine tune the data model according to the processing requirements making it much more reliable and accurate.

The research paper is organized as follows. The second section deals with the review of literature analyzing the existing techniques for pre-processing of social networking data especially twitter feeds. The next section discusses the various pre-processing techniques and the requirements for processing social media data and what kind of output is being expected from such a pre-processing framework and also discusses the importance of stop words and emoticons as a special case with respect to twitter data. Section 4 discusses about the Stanford Twitter Dataset, the Twitter API and provides an overview of obtaining feeds using the Twitter API. Section 5 explains in detail the various parts of the pre-processing framework architecture and about the tools and techniques that are found to be suitable for the same. Section 6 deals with the various platforms and tools that could be employed to handle the pre-processing of twitter data and a discussion about their possible effectiveness from a theoretical stand point. It also deals about the experimental setup and the various parameters considered during the pre-processing phase. Section 7 deals with results and discussion. Section 8 concludes the paper providing a summary of the work and highlighting some future research directions.

Preprocessing

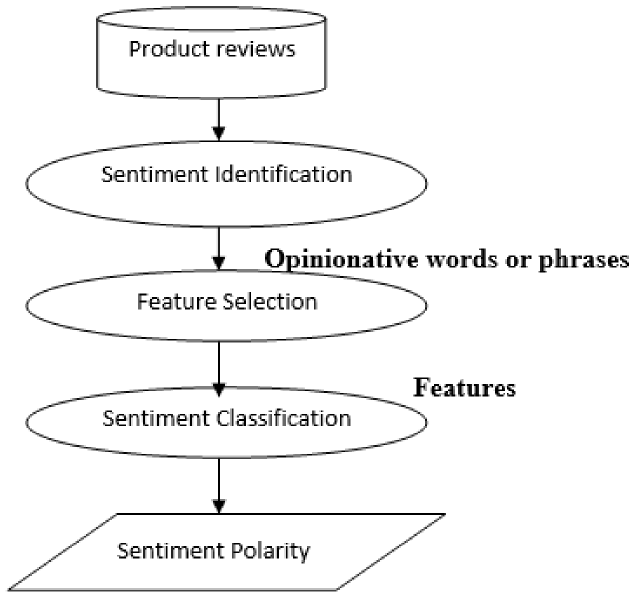
Pre-processing of Twitter data is completely different from the pre-processing done for KDD process in regular text datasets. Tweets are very small by themselves (atmost 140 characters) and because of this users tend to use a different slang or acronyms for almost every common word in English. Usually we eliminate words that are less than 3 characters and greater than 16 characters, but that will not be applicable to tweets. Also most words those are available as part of stop words list may have more importance in a tweet than in a standard text. Say for example the word “but” is very common in English and could be removed as a stop word but for a product review the appearance of this word differentiates between a 4 star and 5 star rating (Say I am satisfied with the product but …). So we will have to determine the uncommonly common words and then only the list of stop words could be decided.

Table 1: List of common positive and negative emoticons used in Twitter

Positive Emoticons	Negative Emoticons
o n n *-* *O* * *	
:P :D :d :p	:(;(:’(;’(
;P ;D ;d ;p	=(={):);
:-) ;-) :-) ;-)	)’:)’;)= }=
:<) :>) ;>) =)	;-{{ ;-{ :-{{ :-{
=} :) (: ;)	:-(;-(:-(;-(
(; :} {: ;}	:,(:’{
{; :]	[: ;]
[; :’) ;’) :-3	
;-3 :-x ;-x :-X	
;-X :-} ;-} :-]	
;-] :-.)	
^_^ ^-^	

Regular text mining approaches eliminates punctuations and symbols but in the case of tweets we have to consider Emoticons, which express sentiment polarity more effectively than words. So this approach cannot remove all the symbols instead it must search for an emoticon based on an In-Memory Dictionary or Hash Table and then decide about removal of symbols if they are not a part of an Emoticon. Emoticons may contain 2 characters, 3 characters or 4 characters as shown in Table 1. So this approach had to concentrate on writing regular expressions to filter them out from a bunch of regular symbols. It also embeds the WordNet corpus to find the Synonyms and Antonyms to effectively identify the words with positive and negative polarity.

Architecture pre-processing framework



The proposed a framework for effective pre-processing of Twitter Feeds is as shown in Figure 1. The framework consists of four major subdivisions from Acquisition of data, followed by two phases of Pre-processing and then Evaluation.

Obtaining Twitter Data

We have used the Streaming Twitter API and the public Streams version of it. All the public data flowing through Twitter currently could be collected using this Public Streams API. Though it is a low latency process it provides considerable amount of data to work on when run continuously for several hours. Twitter Feeds are accessed using the Streaming API and then tweets related to the query passed is obtained and the content is piped in to the next phase of the pre-processing framework.

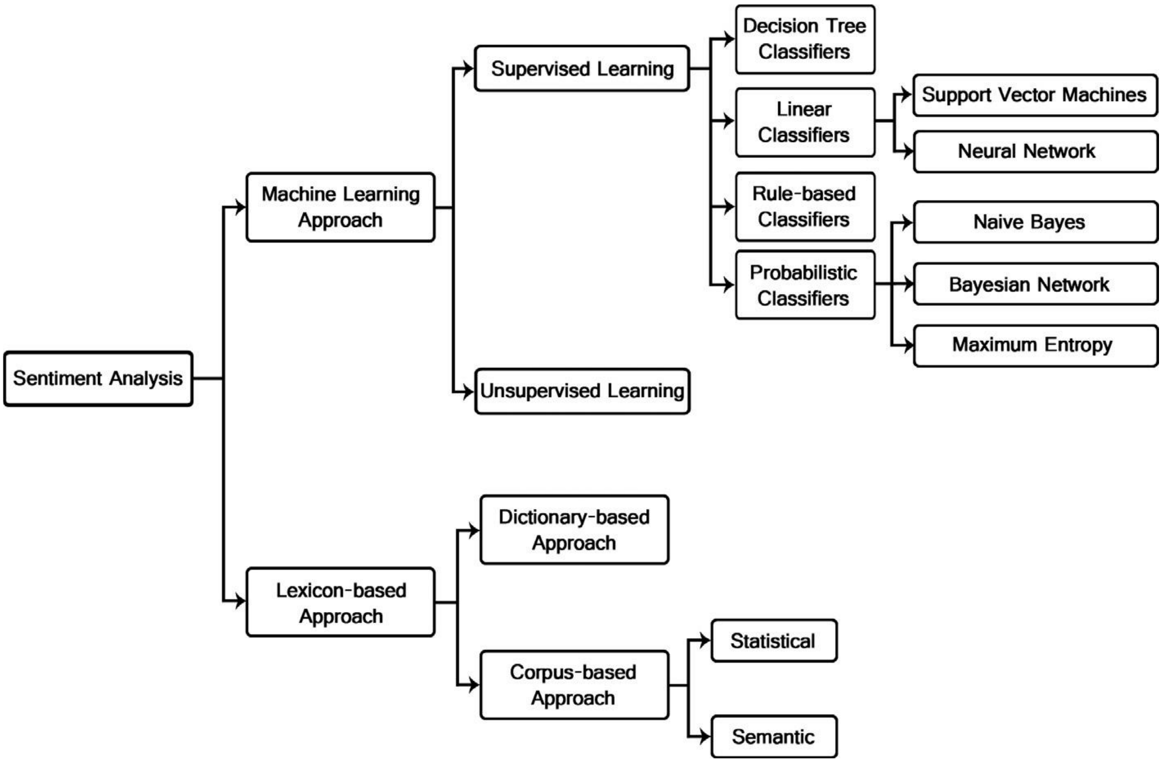
Pre-processing Phase 1

In this phase, the tweets are available as text data and each line contains a tweet. Initially we clean up or remove retweets as that will induce a bias in the classification process. Also we remove the links followed by Hashtags which represents the users. We then process tweets that contain positive emoticons and those that contain negative emoticons. Also there is a possibility of a tweet being neutral. We need to remove the punctuations and other symbols that don’t make any sense as it may result in inefficiencies and may affect the accuracy of the overall process. Also we need to remove tweets that contain both a positive and negative emoticons as it may cause a lot of confusion as far as automatic classification is considered. This results in a clean Twitter Text Corpus which is then sent to the second phase of pre-processing.

Pre-processing Phase 2

In this phase, we tokenize the tweets (split them into individual words of importance). Once tokenized, we perform one of the most important parts of the text pre-processing process, which is normalization. Here we eliminate repeat letters from words that are put in to provide stress or a better context to the tweet. Also we normalize the expressions usually used to express laugh, sorrow etc. which may contain a series of repetitive characters. Then comes the process called lemmatization where the core part of the word alone is considered. It is different from stemming where stemming is just structural, lemmatization is contextual and also takes into account the synonyms and antonyms of a given word. After this process, we remove the stop words as provided by standard stop word corpus. Now the data is ready for evaluation and is sent to the next phase.

Sentiment classification techniques.



Applications of CSIR in social media

There are several applications of context sensitive information retrieval in the social media. Several other possibilities are being explored and commercialized now. With a huge mass of people involved in those activities, almost there is no limit for the amount of personal and private information that is being shared everyday without our knowledge. From job profile matching by Human Resource Recruiters to Home Land Security, everybody is analyzing the social media information for understanding and identifying the required patterns.

Job Profile Matching, Job Search

One of the latest and emerging field and uses of the social media has been Human resource recruitment. Offshore accounts are monitored and their patterns are analyzed and service calls are made to those selective individuals requesting them whether they are interested in switching over to another company in the same country or other. If they are interested they process the request further.

This process is further strengthened by the current social media context as everything including their profile is being shared and also the context sensitive analysis could even suggest their current loyalty level and state of attachment towards the current employer. With such information, the act of pursuing them to switch companies will be a walk in the park. On the other hand the companies can do this kind of analysis in order to identify the potential employees who are likely to leave and then encourage them with incentives and even counselling if necessary.

Problems in Data Handling

The problem with this type of analysis is that the content is going to be in a wide variety of formats and also in various languages. Multilingual context sensitive analysis is one of the research issues and further the various formats in which a resume or profile will be must be taken into consideration. Even the profile information of two persons applying for the same job need not be the same. There could be a different set of documents related to each of their profile.

We need column store technology and also big data processing environment in order to accomplish it. Document databases would be the choice because of its flexibility in storing a wide variety of data. Map-Reduce based processing of information would be the most suitable form of analysis as it is massively parallel and also well suited for this semi-structured data handling.

Conclusion
This paper presents a survey on SA (Sentiment Analysis) techniques that was proposed earlier by researcher. This overview of Sentiment Analysis focuses on Sentiment Polarity implementations, usability and challenges. It also delivers conceptual overview of methodology. Sentiment Analysis in the preprocessing techniques in social media like Twitter, face book, Amazon. Future investigations that are discussed may be implemented in the area of sentiment analysis.

Acknowledgement
I am thankful to Dr.E.Kirubakaran (Research supervisor, Bharathidasan University, Thiruchirapalli.) for his excellent guidance and K.Rajeshwari (Mother), A.Kanagarathinam (Father), K.Gunaseelan (Brother), R.Mohandoss (Spouse) and my children for supporting me in all the ways.

References
1. da Silva, Nádia FF, Eduardo R. Hruschka, and Estevam R. Hruschka., Tweet sentiment analysis with classifier ensembles, Decision Support Systems, 2014.
2. Kranjc, Janez, et al., Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the Clowd Flows platform, Information Processing & Management, 2014.
3. Habernal, Ivan, TomášPtáček, and Josef Steinberger, Supervised sentiment analysis in Czech social media, Information Processing & Management 50.5: 693-707, 2014.
4. Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro., Sentiment analysis in Facebook and its application to e-learning, Computers in Human Behavior 31: 527-541, 2014.
5. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
6. A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, Knowledge and Data Engineering, IEEE Transactions on 23 (3) (2011) 447 462.