REF:JKU/2/83/022

Jomo Kenyatta University of Agriculture and Technology (JKUAT)

School of Computing and Information Technology (SCIT)

Department of Computing

Project Proposal

Title: An Aspect-based Sentiment Analysis technique to predict product performance in promotional campaigns.

Student Name:KIRAGU PHILLIS WACERA Reg. No:CS282-0763/2011

Submission Date:23/03/2015

Supervisor 1:SYLVESTER KIPTOO Sign:_____Date:_____

Supervisor 2:DR.AGNES MINDILA Sign:_____Date:_____

Period: March 2015

Abstract
Textual information in the world can be broadly be categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties. Much of the existing research on textual information processing has been focused on mining and retrieval of factual information, e.g., information retrieval, Web search, text classification, text clustering and many other text mining and natural language processing tasks. Little work had been done on the processing of opinions until only recently. Yet, opinions are so important that whenever we need to make a decision we want to hear others opinions. This is not only true for individuals but also true for organizations.

This document discuss sentiment analysis, the field that is considered with opinion mining from user generated content. It describes the challenges, techniques and application of sentiment analysis. It will concentrate on aspect based analysis which is a sentiment analysis technique that analyses individual phrases in sentences in order to determine the orientation, that is, positive neutral and negative

The result should provide a prototype that is able to take sentiments and determine their orientation and then display the results to the interested parties, that is, promotional campaign managers.

Contents

1 Introduction

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. it is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes,and emotions towards entities such as products, services, organizations,individuals, issues, events, topics, and their attributes.

An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has therefore occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object.

This project research covers techniques and approaches that promise to directly enable the creation of an opinion-oriented information seeking system. The focus is on methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact-based analysis.

The research will involve the use of focused groups and concept analysis methods. The expected outcome is to provide a way to analyze subjective and objective sentences to determine their orientation.

2 Problem Statement

Most of the applications used today to predict results and prices in the market uses fact-based statistics. For example to predict how a product will do in the market after release, it is taken into account statistics such as: number of previously sold items on the same line, amount of revenue gained, percentage of items increase or decrease,etc as much as these statistics are accurate and will help improve products and services offered by that company, it fails to include the crucial information that is obtained from feedback given by the consumers all over the internet and across social media platforms. Therefore the predictions fail to show exactly how people felt about a certain product or service.

Thus understanding the aspect based sentiment analysis technique will help personnel in charge of promotional campaigns to predict results using the peoples opinions and comments

2.1 research objectives

The main objective of this research is to understand aspect based sentiment analysis technique, its challenges and steps so as to identify a way to convert peoples opinions into numerical numbers.

1 Research and identify an algorithm that can be used to implement aspect based sentiment analysis. groups it into either positive, negative or neutral.

2 Identify the needs of the end users who will benefit from the research.

3 Come up with a model that is able to extract data and opinions from various sources and classify them e.g social media and focus groups.

4 Evaluate the model to determine its validity and accuracy through expert review.

2.2 research question

1. Is there an algorithm that is used to implement aspect based sentiment analysis?
2. What are the needs of the end users?
3. How will data and opinions be extracted from the various sources?
4. What is the validity and accuracy of the model and how do experts rate the prototype?

3 Justification

This research will enable promoters to know how their products are fairing in the market. this is because it covers ways to take users feedback as they said it in their natural language and convert it to statistical form i.e numbers. This will help improve the products or promotional methods for consecutive projects.

4 Literature Review

The term sentiment analysis perhaps first appeared in (Nasukawa and Yi,2003), and the term opinion mining first appeared in (Dave, Lawrence and Pennock, 2003). However, the research on sentiments and opinions appeared earlier (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000).(Bing Liu, 2012)

4.1 Techniques involved in Sentiment Analysis

There are many techniques used in sentiment analysis.Researchers in sentiment analysis have focused mainly on two problems: detecting whether the text is subjective or objective, and determining whether the subjective text is positive or negative. The techniques relied on two main approaches: unsupervised sentiment orientation calculation, and supervised and unsupervised classifications based on machine learning. There are four main techniques:(Liu 2012)

4.1.1 Document-Level Sentiment Analysis

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment (Pang, Lee and Vaithyanathan, 2002; Turney, 2002). For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

4.1.2 Sentence-Level Sentiment Analysis

A single document may contain multiple opinions even about the same entities. When there is a need to have a more fine-grained view of the different opinions expressed in the document about the entities we must analyse the sentences within the document. The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification (Wiebe, Bruce and O'Hara, 1999), which distinguishes sentences (called

objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., "I bought this camera last week and the shutter is broken" Researchers have also analyzed clauses (Wilson, Wiebe and Hwa, 2004), but the clause level is still not enough, e.g., "Lenovo is doing very well in this lousy economy."

4.1.3 Comparative Sentiment Analysis

In many cases users do not provide a direct opinion about one product but instead provide comparable opinions such as in these sentences taken from the user forums of Edmonds.com: "300 C touring looks so much better than the Magnum," "I drove the Honda Civic, it does not handle better than the TSX, not even close." The goal of the sentiment analysis system in this case is to identify the sentences that contain comparative opinions, and to extract the preferred entity(ies) in each opinion. One of the pioneering papers on comparative sentiment analysis is Jindal and Liu(2006). This paper found that using a relatively small number of words we can cover 98• Comparative adjectives adverbs such as: 'more,' 'less,' and words ending with –er (for example, 'lighter'). • Superlative adjectives and adverbs such as: 'most,' 'least,' and words ending with –est (for example, 'finest'). • Additional phrases such as 'favor,' 'exceed,' 'outperform,' 'prefer,' 'than,' 'superior,' 'inferior,' 'number one,' 'up against.' Since these words lead to a very high recall, but low precision, a naïve Bayes classifier can be used to filter out sentences that do not contain comparative opinions. The classifier uses sequential patterns as features. The sequential patterns are discovered by the class sequential rule (CSR) mining algorithm. A simple algorithm to identify the preferred entities based on the type of comparative used and the presence of negation is described in Ding et al. (2009).

4.1.4 Aspect-Based Sentiment Analysis

The two previous approaches work well when either the whole document or each individual sentence refers to a single entity. However, in many cases people talk about entities that have many aspects (attributes) and they have a different opinion about each of the aspects. This often happens in reviews about products or in discussion forums dedicated to specific product categories (such as cars, cameras, smartphones, and even pharmaceutical drugs). As an example here is a review of Kindle Fire taken illustrated by Jindal and Liu: "As a long-time Kindle fan I

was eager to get my hands on a Fire. There are some great aspects; the device is quick and for the most part dead-simple to use. The screen is fantastic with good brightness and excellent color, and a very wide viewing angle. But there are some downsides too; the small bezel size makes holding it without inadvertent page-turns difficult, the lack of buttons makes controls harder, the accessible storage memory is limited to just 5GB."

Classifying this review as either positive or negative toward the Kindle would totally miss the valuable information encapsulated in it. The author provides feedback about many aspects of the Kindle (like speed, ease of use, screen quality, bezel size, buttons, and storage memory size). Some of these aspects are reviewed positively while some of the others get a negative sentiment. Aspect-based sentiment analysis (also called feature-based sentiment analysis) is the research problem that focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer.

In this research I will concentrate on the aspect based sentiment analysis. it has two core tasks that hav been extensively researched by many researchers as quoted by Bing Liu in his paper.(liu, 2012). these core tasks are

1. Aspect extraction: This task extracts aspects that have been evaluated.For example, in the sentence, "The image quality of this Camera is amazing," the aspect is "image quality" of the entity represented by "this camera." Note that "this camera" does not indicate the aspect GENERAL here because the evaluation is not about the camera as a whole, but only about its voice quality. However, the sentence "I love this camera" evaluates the camera as a whole, i.e., the GENERAL aspect of the entity represented by "this camera." Bear in mind whenever we talk about an aspect, we must know which entity it belongs to.
2. Aspect sentiment classification: This task determines whether the opinions on different aspects are positive, negative, or neutral. In the first example above, the opinion on the "voice quality" aspect is positive. In the second, the opinion on the aspect GENERAL is also positive.

4.1.5 Aspect extraction

In the context of sentiment analysis, some specific characteristics of the problem can facilitate the extraction. The key characteristic is that an opinion always has a target. The target is often the aspect or topic to be extracted from a sentence. Thus, it is important to recognize each opinion expression and its target from a sentence. However, we should also note that some opinion expressions can play two

roles, i.e., indicating a positive or negative sentiment and implying an (implicit) aspect (target). For example, in "this car is expensive," "expensive" is a sentiment word and also indicates the aspect price. Here, we will focus on explicit aspect extraction. There are four main approaches:

1. Extraction based on frequent nouns and noun phrases.
2. Extraction by exploiting opinion and target relations.
3. Extraction using supervised learning.
4. Extraction using topic modeling.

Since existing research on aspect extraction (more precisely, aspect expression extraction) is mainly carried out in online reviews, we also use the review context to describe these techniques, but there is nothing to prevent them being used on other forms of social media text. There are two common review formats on the Web.

1. Format 1 Pros, Cons, and the detailed review: The reviewer first describes some brief pros and cons separately and then writes a detailed/full review.
2. Format 2 Free format The reviewer writes freely, i.e., no brief pros and cons. Extracting aspects from Pros and Cons in reviews of Format 1 (not the detailed review, which is the same as that in Format 2) is a special case of extracting aspects from the full review and also relatively easy. In (Liu, Hu and Cheng, 2005), a specific method based on a sequential learning method was proposed to extract aspects from Pros and Cons, which also exploited a key characteristic of Pros and Cons, i.e., they are usually very brief, consisting of short phrases or sentence segments. Each segment typically contains only one aspect. Sentence segments can be separated by commas, periods, semi-colons, hyphens, and, but, etc. This observation helps the extraction algorithm to perform more accurately.

4.1.6 sentiment classification

There are two main approaches in determining the orientation of a sentiment, i.e., the supervised learning approach and the lexicon-based approach. However, the key issue is how to determine the scope of each sentiment expression, i.e., whether it covers the aspect of interest in the sentence. The current main approach is to use parsing to determine the dependency and the other relevant information. Supervised learning is dependent on the training data. A model or classifier trained from labeled data in one domain often performs poorly in another domain. Al-

though domain adaptation (or transfer learning) has been studied by researchers (Section 3.4), the technology is still far from mature, and the current methods are also mainly used for document level sentiment classification as documents are long and contain more features for classification than individual sentences or clauses. Thus, supervised learning has difficulty to scale up to a large number of application domains. The lexicon-based approach can avoid some of the issues (Ding, Liu and Yu, 2008; Hu and Liu, 2004), and has been shown to perform quite well in a large number of domains. Such methods are typically unsupervised. They use a sentiment lexicon (which contains a list of sentiment words, phrases, and idioms), composite expressions, rules of opinions (Section 5.2), and (possibly) the sentence parse tree to determine the sentiment orientation on each aspect in a sentence. They also consider sentiment shifters, but-clause and many other constructs which may affect sentiments. Discussed below is the lexicon-based method and it has four steps. Here, we assume that entities and aspects are known:

1. Mark sentiment words and phrases: For each sentence that contains one or more aspects, this step marks all sentiment words and phrases in the sentence. Each positive word is assigned the sentiment score of +1 and each negative word is assigned the sentiment score of 1. For example, the sentence, "The voice quality of this phone is not good, but the battery life is long." After this step, the sentence becomes "The voice quality of this phone is not good [+1], but the battery life is long" because "good" is a positive sentiment word (the aspects in the sentence are italicized). Note that "long" here is not a sentiment word as it does not indicate a positive or negative sentiment by itself in general, but we can infer its sentiment in this context shortly.

2. Apply sentiment shifters: Sentiment shifters (also called valence shifters in (Polanyi and Zaenen, 2004)) are words and phrases that can change sentiment orientations. There are several types of such shifters. Negation words like not, never, none, nobody, nowhere, neither, and cannot are the most common type. This step turns our sentence into "The voice quality of this phone is not good [-1], but the battery life is long" due to the negation word "not."

3. Handle but-clauses: Words or phrases that indicate contrary need special handling because they often change sentiment orientations too. Sentiment Analysis and Opinion Mining 61) The most commonly used contrary word in English is "but". A sentence containing a contrary word or phrase is handled by applying the following rule: the sentiment orientations before the contrary word (e.g., but) and after the contrary word are opposite to each other if the opinion on one side cannot be determined. The if-condition in the rule is

used because contrary words and phrases do not always indicate an opinion change, e.g., "Car-x is great, but Car-y is better." After this step, the above sentence is turned into "The voice quality of this phone is not good [-1], but the battery life is long [+1]" due to "but" ([+1] is added at the end of the but-clause). Notice here, we can infer that "long" is positive for "battery life". Apart from but, phrases such as "with the exception of," "except that," and "except for" also have the meaning of contrary and are handled in the same way. As in the case of negation, not every but means contrary, e.g., "not only ... but also." Such non-but phrases containing "but" also need to be identified beforehand.

4. Aggregate opinions: This step applies an opinion aggregation function to the resulting sentiment scores to determine the final orientation of the sentiment on each aspect in the sentence. This simple algorithm performs quite well in many cases. It is able to handle the sentence "Apple is doing very well in this bad economy" with no problem. Note that there are many other opinion aggregation methods. For example, (Hu and Liu, 2004) simply summed up the sentiment scores of all sentiment words in a sentence or sentence segment. Kim, and Hovy (2004) used multiplication of sentiment scores of words. Similar methods were also employed by other researchers (Wan, 2008; Zhu et al., 2009). To make this method even more effective, we can determine the scope of each individual sentiment word instead of using words distance as above. In Sentiment Analysis and Opinion Mining 62 this case, parsing is needed to find the dependency as in the supervised method discussed above. We can also automatically discover the sentiment orientation of context dependent words such as "long" above. More details will be given in Chapter 6. In fact, the above simple approach can be enhanced in many directions. For example, Blair-Goldensohn et al. (2008) integrated the lexicon-based method with supervised learning. Kessler and Nicolov (2009) experimented with four different strategies of determining the sentiment on each aspect/target (including a ranking method). They also showed several interesting statistics on why it is so hard to link sentiment words to their targets based on a large amount of manually annotated data.

Along with aspect sentiment classification research, researchers also studied the aspect sentiment rating prediction problem which has mostly been done together with aspect extraction in the context of topic modeling.

5 Research Methods and Design

The study will use quantitative data collection methods to get data. This include focused groups and case studies. Having identified the target population as people who are giving feedback to a product or service, the data collection methods identified will be executed as follows:

5.1 Focused groups

This is gathering in-depth information by interviewing six to twelve people in an informal discussion that lasts one to two hours. An experienced interviewer gathers opinions of the group and this is made possible since the opinions are canvassed on specific topics and immediate feedback or additional questions are possible. It can help identify key issues quickly such as understanding why consumers buy or don't buy certain products, identifying the use of products and services. The focus of this study will be on university students who frequently use social media to comment on products.

5.2 Content analysis

Content analysis is a research tool used to determine the presence of certain words or concepts within texts or sets of texts. Researchers quantify and analyze the presence, meanings and relationships of such words and concepts, then make inferences about the messages within the texts, the writer(s), the audience, and even the culture and time of which these are a part. It is often used in quantitative research to study trends or occurrences of information.In collection of data for sentiment analysis, this will be a good method to employ as we will be able to get the data from consumers of a certain product or service based on its time of existence, figure out how much or less it has impacted a given target. To conduct a content analysis on any such text, the text is coded, or broken down, into manageable categories on a variety of levels–word, word sense, phrase, sentence, or theme–and then examined using one of content analysis' basic methods: conceptual analysis or relational analysis.

6 Expected Results

The expected outcome at the end of this study is that there will be an algorithm that will analyse a sentence or statement and extract numerical data that can be used to categorize the entire phrase or sentence as negative or positive. For instance: "This camera has very good image quality but the battery is bad." lets say in the algorithm the phrase "very good" is awarded 6 points and "bad" -3 points, we could then say that the comment is positive with 3 points. a sentence like: "this camera has very good image quality and the battery is ok" If ok has 2 points then the sentence will be positive with 8 points.

7 Schedule

Duration	Description	Deliverables
February-March 2015	Research	Project Proposal
15th March 2015	Presentation	Project Proposal
March - April 2015	Research on area of study	Literature review report
April 2015	Submission	Literature review report
May - June 2015	Analysis and Design System	Design Document
June - July 2015	Coding and Development	System
July 2015	Testing	Test Report
August 2015	Presentation	Project Documentation

8 Budget

•	DESCRIPTION	TOTAL AMOUNT
1	USB Modem	2000
2	Data Bundles	4000
3	Binding and Printing	1000

9 Conclusion

It is clear that people who have an experience with a new or existing product get to express their opinion of it in an unstructured way. This expressions carry sentiments with them that can be analyzed and then used to project the performance

of a product for a target market in a promotional campaign. Thus, this study will provide a platform through which such users can obtain the sentiments, analyze and classify them to determine their orientation that is, whether it's positive, negative or neutral.

References

[1] Liu, b. *Sentiment analysis and opinion mining.*, Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers, 2012.

[2] Turney, P. *Thumbs up or thumbs down?*,Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics, 2002.

[3] Pang, b. and Lee, L. *a sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts.*, In Proceedings of the Association for Computational Linguistics, 2004.

[4] Narayanan, r., Liu, b. and Chaudhary, a. *Sentiment analysis of conditional sentences.*, In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (Singapore), association for Computational Linguistics, 2009.

[5] Tsur, o., Davidov, d. and Rappoport, a. *A great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews.*, In Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[6] Jindal, n. and Liu, b. *identifying comparative sentences in text documents.*, In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval,2014.

[7] Popescu, a.-m. and Etzioni, o. *Extracting product features and opinions from reviews.*, In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2005.

[8] Wu, y., Zhang, q. Huang, X. and Wu, L. *Phrase dependency parsing for opinion mining.*, In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2009.

[9] Ding, X., Liu, b. and Zhang, L. *Entity discovery and assignment for opinion mining applications.*, in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2009.