SENTIMENT ANALYSIS

A Seminar Report

Submitted in partial fulfillment of requirements
for the degree of
Master of Technology (M. Tech)
by
Vaibhav Tripathi
Roll No. 143050009

Under the guidance of
Prof. Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

ABSTRACT

Sentiment analysis is the art of detecting sentiment, opinion or emotion from a piece of text. With the proliferation of social networks, there is now a large amount of opinionated text available on the internet. In recent years, opinionated postings in social media have helped reshape businesses, and sway public sentiments and emotions, impacting our social and political systems. However, sentiment analysis on social media is a challenging task. This report discusses the problems in sentiment analysis in general and sentiment analysis on social media in particular and suggests some solutions.

CONTENTS

1 INTRODUCTION

Natural language processing deals with extracting the subjective information from natural language texts. One such information is the sentiment behind the text. This has also been referred to as opinion/-mood/attitude/appraisal/evaluation in the literature. The simplest form of sentiment analysis involves identifying the polarity of a text, which can be positive, negative or neutral.

An analysis of this form is very useful when the text is large and manual examination is impossible. For instance, a company can easily obtain a feedback on its newly released by performing such an analysis on its review page. Note that this kind of feedback is more natural since the customers/reviewers may write more freely when writing a review in natural language rather than when rating a product numerically (say, on a one to five stars scale) or when filling up feedback forms. More importantly, companies don't have to ask their customers for feedback. The customers post feedbacks at their own leisure in their own writing styles. What is needed is a way to automatically analyse all these reviews and generate some result (may be a numerical score or labels like positive or negative) that may be indicative of the overall customer experience. Sentiment analysis seeks to provide that way.

Putting it simply, if we perform sentiment analysis on a movie review page, and there are more posts like

"Loved the movie! Too good. :)"

and less posts like

"Boring! Wasted time and money.",

we should be able to conclude that the movie review document has a positive sentiment. This is the minimal setting for a sentiment analysis problem.

1.1 SENTIMENT ANALYSIS

Sentiment analysis can be defined as the process of computationally identifying and categorizing sentiments expressed in a piece of text to determine whether the attitude of writer (rather, the sentiment holder) towards a particular topic, product, etc. is positive, negative, or neutral.

1.1.1 *Sentiment analysis research*

Sentiment analysis is a Natural Language Processing (NLP) problem. It touches several aspects of NLP, e.g., coreference resolution, negation handling, and word sense disambiguation, which makes it a difficult task since these are not solved problems in NLP. However, sentiment analysis is a highly restricted NLP problem because the system need not fully understand the semantics the text but only needs to understand some aspects of it, i.e., positive or negative sentiments and their target entities or topics.

Sentiment analysis is a relatively new branch of research, having evolved mainly after 2000 AD. However, the research has taken a serious shape since, and there is a heavy ongoing research in this area. There are several reasons for this.

(a) It has a wide arrange of applications, almost in every domain.

(b) Commercial applications have proliferated.

(c) It offers many challenging research problems, which had never been studied before.

(d) Ever since the evolution of the likes of social networks and blogs, data of unprecedented scale is available for analysis.

Sentiment can be defined as a quintuple:

$$\boxed{(e_i, a_{ij}, s_{ijkl}, h_k, t_l)}$$

Here,

- e_i is the name of an entity
- a_{ij} is an aspect of e_i

- s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i
- h_k is the opinion holder
- t_l is the time when the opinion is expressed by h_k

Thus, the task of setiment analysis can be stated as: Given a sentiment document d, discover all sentiment quintuples $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ in d. However, it is unlikely that we will ever use all of the components of the quintuple in a particular problem.

1.1.2 *Levels of sentiment analysis*

Sentiment analysis can be performed at many levels:

(a) **Document Level:** In this level, we assume that the entire document talks about one particular entity only. In this type of a setting, the opinion holder and time are usually irrelevant. Since we are concerned with just one target throughout the text, the aspect is GENERAL. The opinion tuple thus becomes:

$$\boxed{(e_i, GENERAL, s_{ijkl}, -, -)}$$

(b) **Sentence Level:** This level of analysis goes to the sentences to determine whether each sentence is positive, negative or neutral.

(c) **Aspect Level:** It is based on the idea that an opinion/sentiment consists of a sentiment (positive or negative) and a target. This is a fine-grained analysis where we identify the aspects of the entity being talked about and find the sentiment about each one of them.

For instance, **"In spite of having a great cast, the movie lacks at story."** is positive about the cast but negative about the story.

(d) **User Level:** This is a relatively less discussed level of sentiment analysis. It is different in the sense that in this approach, we don't seek information about the text but the user producing the text. This kind of a problem is more relevant to social media domain since we have more information about the users there (available from the social graph).

1.1.3 *The obvious approach and why it doesn't work*

Not surprisingly, the most important indicators of sentiments are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, good, wonderful, and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Apart from individual words, there could be idioms and phrases that are indicative of a particular sentiment.

The naive approach of sentiment analysis involves identifying sentiment words from the input text, find their polarity scores and them somehow, combine these scores to generate a final sentiment score for the document. Over the years, researchers have designed various sentiment lexicons that contain a list of sentiment words and phrases and their sentiment scores.

While it is established that identifying sentiment words and phrases is crucial to sentiment analysis, using only these for sentiment analysis in a practical setting is far from sufficient. This is due to the follwing issues:

(a) A positive or negative sentiment word may have opposite orientations in different application domains.
For example,

"The plot of the movie is unpredictable"

is positive (keeps the viewers engaged) whereas,

"He is a unpredictable player."

is likely to be negative since we would like players to be consistent and not unpredictable.

(b) A sentence containing sentiment words may not express any sentiment.

The simplest examples of these would be conditional sentences or interrogative sentences. For example,

"Can anyone suggest me a good quality camera?" and **"If you find any fault with this mobile phone, kindly post here."** don't contain sentiment;

whereas,

"Does anyone know how to fix this stupid camera?" and **"If you are planning to buy this phone, please change your plan."** do.

(c) Sarcastic sentences with or without sentiment words are the nightmare of sentiment analyzers. Even humans find it non-trivial to understand sarcasm very often. The main challenge is that in most of the cases, sarcasm detection requires domain-specific world knowledge.

Consider this.

"Had an amazing movie-time. I didn't sleep that good in years!" Now, to actually find out that this review is negative, the system would have to have the knowledge that sleeping during a movie is a bad indicator for the movie. This makes the problem extremely challenging.

(d) Many sentences without sentiment words can also imply opinions. Sentiment can be expressed by objective sentences also, containing just factual information.

"My iphone 6 has bending problems." doesn't contain an opinion as such. Nevertheless, it still has a sentiment.

1.2 APPLICATIONS OF SENTIMENT ANALYSIS

There are interesting applications to sentiment analysis. With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, there is a large amount of user-generated content available to access at all times. Sentiment analysis on this data can yield interesting and useful results. Some of the most common applications of sentiment analysis are:

(a) **Decision Making:** People have always depended on opinions or advice for making important decisions. This was true even before the digital age. But then, you had to reach out to the people you wanted an opinion from, and there were a limited number of people who could be trusted. After that, there was a time of survey forms where a large group of people were expected to voice their opinions and then, the decision would be taken based on the majority. However, people felt bored and tired of filling forms. This was to an extent that the organizations who were interested in those feedbacks started paying the reviewers for filling out the surveys. This was obviously, not a great strategy as the customers were being forced into writing reviews rather than stipulating their own thoughts.

    Now, with the advancement in NLP technologies, organizations can mine the web for relevant data and then analyze it to come up with the information essential for making appropriate decisions. This kind of decision making approach is now being used frequently by producers of goods to identify acceptance of their products by the masses and take production decisions accordingly. Similar approach can also be used to make investment decisions. This is the reason why there have been at least 40-60 start-up companies in the space in the USA alone. Many big corporations have also built their own sentiment systems, e.g., Microsoft, Google, Hewlett-Packard, SAP, and SAS.

(b) **Prediction:** Here is the most interesting application of sentiment analyzers. Public inclination is responsible for a lot of important happenings. This ranges from political elections to box office collections. If we can understand the sentiment of the masses, we may be able to predict successfully the outcome of such public-influenced events.

    Sentiment analyzers have been used to successfully predict electoral results, stock markets and movie revenues in the past. Some researchers have also used the setiments behind the text to predict the gender of the person generating the text. A more recent example comes from the UK, where Tweview attempted to predict the winner of the 2013 X Factor using sentiment, volume, and a lot of other social factors. While they predicted Nicholas McDonald as the winner of 2013 X Factor, he ended up second, with Sam Bailey winning. However, the predictions made all throughout the show were very close to the actual results, revealing the potential of sentiment systems.

1.3 THWARTING - A PARTICULARLY IMPORTANT PROBLEM

Thwarting is defined by Pang et al., (2008) as follows:

Thwarted expectations basically refer to the phenomenon wherein the author of the text first builds up certain expectations for the topic, only to produce a deliberate contrast to the earlier discussion.

Consider this example:

"The movie didn't have a story at all. The music was dull. Also, the direction could have been a lot better. But I lovvved the movie because Caprio is in it! I love Caprio!"

Notice that the above movie review is negative for most part. But as we move to the last segment of the sentence, there is a shift in polarity from negative to positive and the positive dominates. So, if we are not dealing with the aspect level analysis, the overall polarity of this review must be positive.

For computational purpose, thwarting is defined as:

The phenomenon wherein the overall polarity of the document is in contrast with the polarity of majority of the document.

The problem at hand is detecting the correct polarity of such thwarted documents. This is easy if we can detect somehow whether the document is thwarted or not (then we can deal with it accordingly). This leads to a simpler problem which is to identify whether a given document is thwarted or not. This is called as the problem of detecting turnarounds in sentiment.

To handle this problem, an aspect level of analysis is required. If we treat out target as a combination of subtargets arranged in a hierarchical fashion (with the original entity as root) , thwarting can be considered as the phenomenon of polarity reversal at a higher level of the hierarchy compared to the polarity at lower level. Thus, the proposed solution starts with constructing a domain ontology of the system under consideration.

For example, if we are detecting turnarounds in movie reviews, we will have to find the ontology of a movie. This can be done manually as well as automated, using techniques like Latent Dirichlet Allocation (LDA). LDA can learn the key features of a movie (e.g. cast, acting, music etc.) from a movie review corpus. If LDA misses out on some features, human intervention may be required to provide the necessary components. The obtained list of all components should be then arranged into a proper hierarchy by a human annotator.

The naive approach to detect thwarting after this would be to check the sentiment score obtained at all nodes in the ontology tree. If different levels exhibit different polarity, the document is likely to be thwarted. In procedure, this is done in three distinct steps:

(a) **Dependency parse:** For every input document, it has to undergo a dependency parsing first. This is essential to identify all the adjective-noun relations.

(b) **Obtaining polarity at nodes:** Now, for all extracted nouns, we check if it is present in the domain ontology. If it is, the score for the corresponding node will be the determined by the associated adjective word.

(c) **Thwarting detection:** Now that we have sentiment scores for all nodes in the ontology, we can summarize the sentiment polarity at each level of the ontology. If there is a polarity reversal between levels, we can judge the document as thwarted.

However, this method fails to perform well in practical situations, since nodes at the same level may not be equally important for the sentiment of the document. For example, For example, say in a camera ontology, the body and video capability might be subjective whereas any fault in the lens or the battery will render the camera useless, hence they are more critical. Therefore, a relative weighing is required between all features of the ontology.

The idea used here is of percolated polarities. This means that polarity at a node is its polarity in the document along with the polarities of all its descendants. This is based on the intuition that every word contributes some polarity to its parent node in the domain ontology.

We can also define controlled percolation, wherein the value added for a particular descendant is a function of its distance from the node.

For example,

$$p(movie) = p(movie) + p(cast)/2 + p(plot)/2 + p(music)/2 + p(acting)/4 + p(script)/4 + p(cinematography)/4$$

A more complicated approach involves using the above as preprocessing to obtain features and then pose the problem of thwarting detection as a classification problem. Ramteke et al. (2013) have used SVM classifier on such features to achieve an AUC score as high as 81%.

2 EXAMPLES

Ei choro aeterno antiopam mea, labitur bonorum pri no **(author?)** [1]. His no decore nemore graecis. In eos meis nominavi, liber soluta vim cu. Sea commune suavitate interpretaris eu, vix eu libris efficiantur.

2.1 A NEW SECTION

Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

Examples: *Italics*, ALL CAPS, Small Caps, low small caps.

2.1.1 *Test for a Subsection*

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Note: The content of this chapter is just some dummy text. It is not a real language.

Errem omnium ea per, pro **UML!** (**UML!**) congue populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

Ei choro aeterno antiopam mea, labitur bonorum pri no. His no decore nemore graecis. In eos meis nominavi, liber soluta vim cu.

2.1.2 *Autem Timeam*

Nulla fastidii ea ius, exerci suscipit instructior te nam, in ullum postulant quo. Congue quaestio philosophia his at, sea odio autem vulputate ex. Cu usu mucius iisque voluptua. Sit maiorum propriae at, ea cum **API!** (**API!**) primis intellegat. Hinc cotidieque reprehendunt eu nec. Autem timeam deleniti usu id, in nec nibh altera.

Non vices medical da. Se qui peano distinguer demonstrate, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primarimente su del.[1]

Sia ma sine svedese americas. Asia **(author?)** [2] representantes un nos, un altere membros qui.[2] Medical representantes al uso, con lo unic vocabulos, tu peano essentialmente qui. Lo malo laborava anteriormente uso.

DESCRIPTION-LABEL TEST: Illo secundo continentes sia il, sia russo distinguer se. Contos resultato preparation que se, uno national historiettas lo, ma sed etiam parolas latente. Ma unic quales sia. Pan in patre altere summario, le pro latino resultato.

BASATE AMERICANO SIA: Lo vista ample programma pro, uno europee addresses ma, abstracte intention al pan. Nos duce infra publicava le. Es que historia encyclopedia, sed terra celos avantiate in. Su pro effortio appellate, o.

Tu uno veni americano sanctificate. Pan e union linguistic **(author?)** [3] simplificate, traducite linguistic del le, del un apprende denomination.

2.2.1 *Personas Initialmente*

Uno pote summario methodicamente al, uso debe nomina hereditage ma. Iala rapide ha del, ma nos esser parlar. Maximo dictionario sed al.

2.2.1.1 *A Subsubsection*

Deler utilitate methodicamente con se. Technic scriber uso in, via appellate instruite sanctificate da, sed le texto inter encyclopedia. Ha iste americas que, qui ma tempore capital.

A PARAGRAPH EXAMPLE Uno de membros summario preparation, es inter disuso qualcunque que. Del hodie philologos occidental al, como publicate litteratura in web. Veni americano **(author?)** [4] es con, non internet millennios secundarimente ha. Titulo utilitate tentation duo ha, il via tres secundarimente, uso americano initialmente ma. De duo deler personas initialmente. Se duce facite westeuropee web, Table 1 nos clave articulos ha.

[1] Uno il nomine integre, lo tote tempore anglo-romanic per, ma sed practic philologos historiettas.

[2] De web nostre historia angloromanic.

LABITUR BONORUM PRI NO	QUE VISTA	HUMAN
fastidii ea ius	germano	demonstratea
suscipit instructior	titulo	personas
quaestio philosophia	facto	demonstrated **(author?)**

Table 1: Autem timeam deleniti usu id. **(author?)**

A. Enumeration with small caps (alpha)

B. Second item

Medio integre lo per, non **(author?)** [5] es linguas integre. Al web altere integre periodicos, in nos hodie basate. Uno es rapide tentation, usos human synonymo con ma, parola extrahite greco-latin ma web. Veni signo rapide nos da.

2.2.2 *Linguistic Registrate*

Veni introduction es pro, qui finalmente demonstrate il. E tamben anglese programma uno. Sed le debitas demonstrate. Non russo existe o, facite linguistic registrate se nos. Gymnasios, e. g., sanctificate sia le, publicate Figure 1 methodicamente e qui.

Lo sed apprende instruite. Que altere responder su, pan ma, i. e., signo studio. Figure 1b Instruite preparation le duo, asia altere tentation web su. Via unic facto rapide de, iste questiones methodicamente o uno, nos al.

(a) Asia personas duo.

(b) Pan ma signo.

(c) Methodicamente o uno.

(d) Titulo debitas.

Figure 1: Tu duo titulo debitas latente.

CONCLUSIONS

Sentiment analysis is an emerging field with a thrilling potential. It, along with the power of social networks, can harness a great deal of unseen and useful information. Since this information is coming from people, it allows the providers to be in touch with their consumers at all times, understanding their demands and grievances. Sentiment analysis systems, thus, seek to create a world where the voice of the people is given a status as never before and their needs are always understood, wherever and howsoever, they may express it.

It being established that sentiment analysis systems can achieve a lot, some doubt the future of research in this direction, saying that sentiment systems can never achieve an accuracy at which they can reliable enough to be instrumental to global change. The realization to be made here is that the gold standard here is not a full 100%.

Sentiment is humanly. So, the performance of sentiment analysis systems has to be measured against a human's ability to do the same. Incidentally, there are studies which show that there is an agreement of about 70% among humans regarding sentiment on any subject. This is called as human concordance. So, this is the accuracy that sentiment systems target, which is very much achievable.

BIBLIOGRAPHY

[1] G. Dueck, *Dueck's Trilogie: Omnisophie – Supramanie – Topothesie*. Springer, Berlin, 2005. `http://www.omnisophie.com`.

[2] J. Bentley, *Programming Pearls*. Boston, MA, USA: Addison–Wesley, 2nd ed., 1999.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: The MIT Press, 2nd ed., 2001.

[4] D. E. Knuth, "Big Omicron and Big Omega and Big Theta," *SIGACT News*, vol. 8, pp. 18–24, April/June 1976.

[5] I. Sommerville, *Software Engineering*. Boston, MA, USA: Addison-Wesley, 4th ed., 1992.