Analytical Procedure for Sentiment Analysis in Opinion Mining

[1]Appireddy V Nagireddy, Asst.prof, KMIT, Narayanaguda, a.v.nagireddy@gmail.com
[2]Nellore Sudha, Asst.prof, KMIT, Narayanaguda, kmitsudhakar@gmail.com
[3]Kethineni Venkateswarlu, Asst.prof, KMIT, Narayanaguda, venkat0534@gmail.com

***Abstract*—Opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is not only true for individuals but also true for organizations**

Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining.sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. Which is often considered the unstructured data, this book takes a structured approach introducing the problem with the aim of bridging the unstructured and structured worlds and facilitating qualitative and quantitative analysis of opinions. This is crucial for practical applications.I first define the problem in order to provide an abstraction or structure to the problem. The abstraction, we will naturally see its key sub- problems. The subsequent chapters discuss the existing techniques for solving these sub problems.

I. Introduction

1.1 Sentiment Analysis: A Fascinating Problem

Sentiment analysis, also called opinion mining, 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. It rep- resents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study. The term sentiment analysis perhaps first appeared in (Nasukawa and Yi, 2003), and the term opinion mining first appeared.

Although linguistics and natural language processing (NLP) have a long history, little research had been done about people's opinions and sentiments before the year 2000. Since then, the field has become a very active research area. There are several reasons for this. First, it has a wide arrange of applications, almost in every domain.

1.2 Sentiment Analysis Applications

Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make adecision, we want to know others' opinions.

In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies. 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, individuals and organizations are increasingly using the content in these media for decision-making.

Sentiment analysis applications have spread to almost every possible domain, from consumer products, services, healthcare, and financial services to social events and political elections. I myself have implemented a sentiment analysis system called Opinion Parser, and worked on projects in all these areas in a start-up company. 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 in-house capabilities, e.g., Microsoft, Google, Hewlett- Packard, SAP, and SAS. These practical applications and industrial interests have provided strong motivations for research in sentiment analysis.

1.3 Natural Language Processing Issues

Finally, we must not forget sentiment analysis is a NLP problem. It touches every aspect of NLP, e.g., co reference resolution, negation handling, and word sense disambiguation, which add more difficulties since these are not solved problems in NLP.

However, it is also useful to realize that sentiment analysis is a highly restricted NLP problem because the system does not need to fully understand the semantics of each sentence or document but only needs to understand some aspects of it, i.e., positive or negative sentiments and their target entities or topics. In this sense, sentiment analysis offers a great platform for NLP researchers to make tangible progresses on all fronts of NLP with the potential of making a huge practical impact. NLP to join force to make a concerted effort to solve the problem. commercial companies that are in the business of writing fake reviews and bogus blogs for their clients.

Several high profile cases of fake reviews have been reported in the news. It is important to detect such spamming activities to ensure that the opinions on the Web are a trusted source of valuable in- formation. Unlike extraction of positive and negative opinions, opinion spam detection is not just an NLP problem as it involves the analysis of people's posting behaviors. It is thus also a data mining problem.

1.4 Types of Opinions

The type of opinions that we have discussed so far is called *regular opinion* another type is called *comparative opinion*. In fact, we can also classify opinions based on how they are expressed in text, *explicit opinion* and *implicit* **(**or *implied***)** *opinion*.

1.4.1 Regular and Comparative Opinions

A *regular opinion* is often referred to simply as an *opinion* in the literature and it has two main sub-types

Direct opinion: A *direct opinion* refers to an opinion expressed directly on an entity or an entity aspect, e.g., "*The picture quality is great*."

Indirect opinion: An *indirect opinion* is an opinion that is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. This sub-type often occurs in the medical domain. for example, the sentence "*After injection of the drug, my joints felt worse*" describes an undesirable effect of the drug on "my joints", which indirectly gives a negative opinion or sentiment to the drug. In the case, the entity is the *drug* and the aspect are the *effect on joints*.

A *comparative opinion* expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities (Jindal and Liu, 2006a; Jindal and Liu, 2006b). For example, the sentences, "*Coke tastes better than Pepsi*" and "*Coke tastes the best*" express two comparative opinions.

2. Existing System

The problem definitions state what kind of summary may be desired. Along with the problem definitions, the chapter will also discuss several related concepts such as subjectivity and emotion.

I mainly use reviews and sentences from reviews as examples to introduce ideas and to define key concepts, but the ideas and the resulting definitions are general and applicable to all forms of formal and informal opinion text such as news articles, tweets (Twitter postings), forum discussions, blogs, and Face book postings. We use the following review about a Canon camera to introduce the problem

(1) *I bought a Canon G12 camera six months ago*.
(2) *I simply love it*.
(3) *The picture quality is amazing*.
(4) *The battery life is also long*.
(5) *However, my wife thinks it is too heavy for her*.

(1) The review has a number of opinions, both positive and negative, about Canon G12 camera. Sentence (2) expresses a positive opinion about the Canon camera as a whole. Sentence (3) expresses a positive opinion about its picture equality. Sentence (4) expresses a positive opinion about its battery life. Sentence (5) expresses a negative opinion about the weight of the camera.

Observation: An opinion consists of two key components: a target g and a sentiment s on the target This review has opinions from two persons, which are called *opinion sources* or *opinion holders*
The holder of the opinions in sentences (2), (3), and (4) is the author of the review ("John Smith"), but for sentence (5), it is the wife of the author. The date of the review is September 10, 2017. This date is important in practice because one often wants to know how opinions change with time and opinion trends.

Definition (Opinion): An *opinion* is a quadruple. (g, s, h, t), where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed.

This definition, although quite concise, may not be easy to use in practice especially in the domain of online reviews of products, services, and brands because the full description of the target can be complex and may not even appear in the same sentence.

For example, in sentence (3), the opinion target is actually "picture quality of Canon G12", but the sentence mentioned only "picture quality". In this case, the opinion target is not just "picture quality" because without knowing that the sentence is evaluating the picture quality of the Canon G12 camera, the opinion in sentence (3) alone is of little use. In practice, the target can often be decomposed and described in a structured manner with multiple levels,

3. Proposed System

3.1 Sentiment Classification Using Supervised Learning

Sentiment classification is usually formulated as a two-class classification problem, *positive* and *negative*. Training and testing data used are normally product reviews. Since online reviews have rating scores assigned by their reviewers, e.g., 1-5 stars, the positive and negative classes are determined using the ratings. For example, a review with 4 or 5 stars is considered a positive review, and a review with 1 to 2 stars is considered a negative review. Most research papers do not use the neutral class, which makes the classification problem considerably easier, but it is possible to use the neutral class, e.g., assigning all 3-star reviews the neutral class.

Sentiment classification is essentially a text classification problem. Traditional text classification mainly classifies documents of different topics, e.g., politics, sciences, and sports. In such classifications, topic related words are the key features. However, in sentiment classification, sentiment or opinion words that indicate positive or negative opinions are more important, e.g., *great*, *excellent*, *amazing*, *horrible*, *bad*, *worst*, etc.

Part of speech. The part-of-speech (POS) of each word can be important too. Words of different parts of speech (POS) may be treated differently. For example, it was shown that adjectives are important indicators of opinions. Thus, some researchers treated adjectives as special features. However, one can also use all POS tags and their n-grams as features. Note that in this book, we use the standard *Penn Treebank POS Tags* as shown in Table 3.1 The Penn Treebank site is at http://www.cis.upenn.edu/ treebank/ home.html.

Sentiment words are words in a language that are 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. Most sentiment words are adjectives and adverbs, but nouns (e.g., *rubbish*, *junk*, and *crap*) and verbs (e.g., *hate* and *love*) can also be used to express sentiments. Apart from individual words, there are also *sentiment phrases* and *idioms*, e.g., *cost someone an arm and a leg*.

Rules of opinions. Apart from sentiment words and phrases, there are also many other expressions or language compositions that can be used to express or imply sentiments and opinions. *Sentiment shifters* These are expressions that are used to change the sentiment orientations, e.g., from positive to negative or vice versa. Negation words are the most important class of sentiment shifters. For example, the sentence "*I don't like this camera*" is negative. There are also several other types of sentiment shifters. We will discuss them shifters also need to be handled with care because not all occurrences of such words mean sentiment changes. For example, "not" in "not only … but also" does not change sentiment orientation.

Syntactic dependency. Words dependency-based features generated from parsing or dependency trees are also tried by researchers.

Table 3.1: Penn Treebank Part-of-Speech (POS) tags

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Preposition or subordinating conjunction	SYM	Symbol
IN	Adjective	TO	*To*
JJ	Adjective, comparative	UH	Interjection
JJR	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
MMS	Noun, plural	VBZ	Verb, non-3rd person singular present

3.2 Sentiment Classification Using Unsupervised Learning

Since sentiment words are often the dominating factor for sentiment classification, it is not hard to imagine that sentiment words and phrases may be used for sentiment classification in an unsupervised manner.

The method in such a technique. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions. The syntactic patterns are composed based on part-of-speech (POS) tags.
The algorithm given in consists of three steps:
Step 1: Two consecutive words are extracted if their POS tags conform to any of the patterns in Table 3.2. For example, pattern 2 means that two consecutive words are extracted if the first word is an adverb, the second word is an adjective, and the third word (not extracted) is not a noun. As an example, in the sentence "*This piano produces beautiful sounds*", "*beautiful sounds*" is extracted as it satisfies the first pattern. The reason these patterns are used is that JJ, RB, RBR and RBS words often express opinions. The nouns or verbs act as the contexts because in different contexts a JJ, RB, RBR and RBS word may express different sentiments.
For example, the adjective (JJ) "unpredictable" may have a negative sentiment in a car review as in "unpredictable steering," but it could have a positive sentiment in a movie review as in "unpredictable plot."
Step 2: It estimates the sentiment orientation (*SO*) of the extracted phrases using the *point wise mutual information* (PMI) measure:

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right)$$

PMI measures the degree of statistical dependence between two terms.
Here, Pr(*term*1 ^ *term*2) is the actual co-occurrence probability of *term*1 and *term*2, and Pr(*term*1)Pr(*term*2) is the co-occurrence probability of the two terms if they are statistically independent. The sentiment orientation (*SO*) of a phrase is computed based on its association with the positive reference word "excellent" and the negative reference word "poor":

$$SO(phrase) = PMI(phrase, \text{"excellent"}) - PMI(phrase, \text{"poor"}).$$

The probabilities are calculated by issuing queries to a search engine and collecting the number of *hits*. For each search query, a search engine usually gives the number of relevant documents to the query, which is the number of hits. Thus, by searching the two terms together and separately.

	First word	Second word	Third word
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table: Patterns of POS tags for extracting two-word phases

The probabilities in Equation (1) can be estimated. AltaVista search engine was used because it has a NEAR operator to constrain the search to documents that contain the words within ten words of one another in either order. Let h*its* (*query*) be the number of hits returned.
Equation can be rewritten as:

$$SO(phrase) = \log_2\left(\frac{hits(\text{phrase } NEAR \text{ "excellent"})hits(\text{"poor"})}{hits(\text{phrase } NEAR \text{ "poor"})hits(\text{"excellent"})}\right)$$

Step 3: Given a review, the algorithm computes the average *SO* of all phrases in the review and classifies the review as positive if the average *SO* is positive and negative otherwise.

Final classification accuracies on reviews from various domains range from 84% for automobile reviews to 66% for movie reviews.

Another unsupervised approach is the lexicon-based method, which uses a dictionary of sentiment words and phrases with their associated orientations and strength, and incorporates intensification and negation to compute a sentiment score for each document.

4. Conclusion and Future enhancements

This paper introduced the field of sentiment analysis and opinion mining and surveyed the current state-of-the-art. The existing techniques for dealing with them were discussed. After that, the book discussed the problem of sentiment lexicon generation. Two dominant approaches were covered. Completely automated and accurate solution is nowhere in sight. However, it is possible to devise effective semi-automated solutions. The key is to fully understand the whole range of issues and pitfalls, cleverly manage them, and determine what portions can be done automatically and what portions need human assistance. In the continuum between the fully manual solution and the

fully automated solution, as time goes by, we can push more and more towards automation. I do not see a silver bullet solution soon. A good bet would be to work hard on a large number of diverse application domains, understand each of them, and design a general solution gradually.

References

[1] Andreevskaia, Alina and Sabine Bergler. *When specialists and generalists work together: Overcoming domain dependence in sentiment tagging*. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2008)*. 2008.

[2] Andrzejewski, David and Xiaojin Zhu. *Latent Dirichlet Allocation with topic-in-set knowledge*. in *Proceedings of NAACL HLT*. 2009.

[3] Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. *Show me the money!: deriving the pricing power of product features by mining consumer reviews*. in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2007)*. 2007.

[4] Banea, Carmen, Rada Mihalcea, and Janyce Wiebe. *Multilingual subjectivity: are more languages better?* in *Proceedings of the International Conference on Computational Linguistics (COLING-2010)*. 2010.

[5] Barbosa, Luciano and Junlan Feng. *Robust sentiment detection on twitter from biased and noisy data*. in *Proceedings of the International Conference on Computational Linguistics (COLING-2010)*. 2010.

[6] Benamara, Farah, Baptiste Chardon, Yannick Mathieu, and Vladimir Popescu. *Towards Context-Based Subjectivity Analysis*. in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2011)*. 2011.

[7] Burfoot, Clinton, Steven Bird, and Timothy Baldwin. *Collective classification of congressional floor-debate transcripts*. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*. 2011.

[8] Carvalho, Paula, Luís Sarmento, Jorge Teixeira, and Mário J. Silva. *Liars and saviors in a sentiment annotated corpus of comments to political debates*. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*. 2011.

[9] Choi, Yejin and Claire Cardie. *Hierarchical sequential learning for extracting opinions and their attributes*. in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010)*. 2010.

[10] Das, Dipanjan. *A Survey on Automatic Text Summarization Single- Document Summarization.* Language, 2007. 4: p. 1-31.

[11] Davidov, Dmitry, Oren Tsur, and Ari Rappoport. *Enhanced sentiment learning using twitter hashtags and smileys*. in *Proceedings of Coling- 2010*. 2010.

[12] Feldman, Ronen, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. *The Stock Sonar - Sentiment Analysis of Stocks Based on a Hybrid Approach*. in *Proceedings of 23rd IAAI Conference on Artificial Intelligence (IAAI-2011)*. 2011.

[13] Goldberg, Andrew B. and Xiaojin Zhu. *Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization*. in *Proceedings of HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*. 2006.

[14] Hardisty, Eric A., Jordan Boyd-Graber, and Philip Resnik. *Modeling perspective using adaptor grammars*. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)* 2010.