International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Sentiment Analysis Twitter Data - An Overview

Saumya Shukla
Computer Science & Engg. Department
AIET, Lucknow, India
saumyashukla2307@gmail.com

Manmohan Singh Yadav
Computer Science & Engg. Department
AIET, Lucknow, India
yadavmohanman100@gmail.com

***Abstract*--Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others' opinions. Enveloping real-life applications are only part of the reason why sentiment analysis is a popular research problem. It is also highly challenging as a NLP research topic, and covers many novel sub problems as we will see later. Additionally, there was little research before the year 2000 in either NLP or in linguistics. Part of the reason is that before then there was little opinion text available in digital forms. Since the year 2000, the field has grown rapidly to become one of the most active research areas in NLP.**

Key words : SA, Opinion Mining, NPL, Data Mining

1 Introduction

In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers additionally want to understand the opinions of existing users of a product before buying it, and others' opinions about political candidates before making a option call in a political election. In the past, when an individual required opinions, he/she asked friends and family. When an organization or a business required public or shopper opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and shopper opinions has long been a vast business itself for selling, 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 more and more exploitation the content in these media for decision making. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may now not be necessary to conduct surveys, opinion polls, and focus groups in order to collect public opinions as a result of there's an abundance of such information publically accessible. However, finding and monitoring opinion sites on the internet and distilling the data contained in them remains a formidable task owing to the proliferation of numerous sites. Each web site generally contains a immense volume of opinion text that's not invariably simply deciphered in long blogs and forum postings. The average human reader will have issue distinguishing relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are therefore required. In recent years, we have witnessed that opinionated postings in social media have helped reshape businesses, and sway public sentiments and emotions, which have deeply compact on our social and political systems.. It has thus become a necessity to gather and study opinions on the online. Of course, opinionated documents not solely exist on the internet (called external data), many organizations additionally have their internal data, e.g., customer feedback collected from emails and call centers or results from surveys conducted by the organizations. Due to these applications, industrial activities have flourished in recent years. Sentiment analysis applications have spread to nearly every potential domain, from consumer merchandise, services, healthcare, and financial services to social events and political elections.

2. Different Levels of Analysis

We now give a brief introduction to the main research problems based on the level of granularities of the existing research. In general, sentiment analysis has been investigated mainly at two levels:

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion regarding the product. This task is commonly called document-level sentiment classification. This level of analysis assumes that every document expresses opinions on one entity (e.g., a single product). Thus, it is not applicable to documents which assess or compare multiple entities.

Sentence level: The task at this level goes to the sentences and determines whether every sentence expressed a positive, negative, or neutral opinion. Neutral usually means that no opinion. This level of analysis is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that categorical factual data from sentences (called subjective sentences) that categorical subjective views and opinions.

3. Opinion Summarization

Unlike factual information, opinions are essentially subjective. One opinion from a single opinion holder is usually not sufficient for action. In most applications, one needs to analyze opinions from a large number of people. This indicates that some form of summary of opinions is desired. Although an opinion summary can be in one of many forms, e.g., structured summary

(see below) or short text summary, the key components of a summary should include opinions about different entities and their aspects and should also have a quantitative perspective. The quantitative perspective is especially important because 20% of the people being positive about a product is very different from 80% of the people being positive about the product. The opinion quintuple defined above actually provides a good source of information and also a framework for generating both *qualitative* and *quantitative* summaries. A common form of summary is based on aspects and is called *aspect-based opinion summary* (or *feature-based opinion summary)*

4. Different Types of Opinions

The type of opinions that we have discussed so far is called *regular opinion* (Liu, 2006 and 2011). 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*.

4.1 Regular and Comparative Opinions Regular opinion:

A *regular opinion* is often referred to simply as an *opinion* in the literature and it has two main sub-types (Liu, 2006 and 2011):

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 is the *effect on joints*. Much of the current research focuses on direct opinions. They are simpler to handle. Indirect opinions are often harder to deal with. For example, in the drug domain, one needs to know whether some desirable and undesirable state is before or after using the drug. For example, the sentence "*Since my joints were painful, my doctor put me on this drug*" does not express a sentiment or opinion on the drug because "painful joints" (which is negative) happened before using the drug.

Comparative opinion: 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. A comparative opinion is usually expressed using the *comparative* or *superlative* form of an adjective or adverb, although not always (e.g., *prefer*). Comparative opinions also have many types.

4.2 Explicit and Implicit Opinions

Explicit opinion: An *explicit opinion* is a subjective statement that gives a regular or comparative opinion, e.g.,
"*Coke tastes great*," and
"*Coke tastes better than Pepsi*."

Implicit (or implied) opinion: An *implicit opinion* is an objective statement that implies a regular or comparative opinion. Such an objective statement usually expresses a desirable or undesirable fact, e.g., "*I bought the mattress a week ago, and a valley has formed*," and "*The battery life of Nokia phones is longer than Samsung phones*." Explicit opinions are easier to detect and to classify than implicit opinions. Much of the current research has focused on explicit opinions. Relatively less work has been done on implicit opinions. In a slightly different direction, studied the influence of syntactic choices on perceptions of implicit sentiment. For example, for the same story, different headlines can imply different sentiments.

5. Subjectivity and Emotion

There are two vital ideas that are closely connected to sentiment and opinion, i.e., subjectivity and emotion.
Definition (sentence subjectivity): An objective sentence presents some factual information concerning the world, while a subjective sentence expresses some personal feelings, views, or beliefs. An example objective sentence is "iPhone is an Apple product." An example subjective sentence is "I like iPhone." Subjective expressions come in several forms, e.g., opinions, allegations, desires, beliefs, suspicions, and speculations. There is some confusion among researchers to equate subjectivity with opinionative. By opinionated, we mean that a document or sentence expresses or implies a positive or negative sentiment. The two ideas aren't equivalent, although they have an outsized intersection. The task of determining whether or not a sentence is subjective or objective is referred to as subjectivity classification. Here, we ought to note the following:

- A subjective sentence may not categorical any sentiment. For example, "I think that he went home" is a subjective sentence, but will not specific any sentiment. Sentence (5) in Example 4 is additionally subjective however it doesn't provides a positive or negative sentiment concerning anything.
- Objective sentences can imply opinions or sentiments due to desirable and undesirable facts (Zhang and Liu, 2011b). For example, the following two sentences that state some facts clearly imply negative sentiments (which are implicit

opinions) concerning their individual merchandise as a result of the facts are undesirable:

"The earphone broke in two days."

"I brought the mattress a week past and a natural depression has formed"

Apart from explicit opinion bearing subjective expressions, many other types of subjectivity have also been studied although not as extensive, e.g., affect, judgment, appreciation, speculation, hedge, perspective, arguing, agreement and disagreement, political stances Many of them may also imply sentiments.

Definition (emotion): Emotions are our subjective feelings and thoughts. Emotions have been studied in multiple fields, e.g., psychology, philosophy, and sociology. The studies are very broad, from emotional responses of physiological reactions (e.g., heart rate changes, blood pressure, sweating and so on), facial expressions, gestures and postures to different types of

subjective experiences of an individual's state of mind. Scientists have categorized people's emotions into some categories. However, there is still not a set of agreed basic emotions among researchers. Based on (Parrott, 2001), people have six primary emotions, i.e., *love*, *joy*, *surprise*, *anger*, *sadness*, and *fear*, which can be sub-divided into many secondary and tertiary emotions. Each emotion can also have different intensities. Emotions are closely related to sentiments. The strength of a sentiment or opinion is typically linked to the intensity of certain emotions, e.g., *joy* and *anger*. Opinions that we study in sentiment analysis are mostly *evaluations* (although not always). According to consumer behavior research, evaluations can be broadly categorized into two types: *rational evaluations* and *emotional evaluations* (Chaudhuri, 2006).

Rational evaluation: Such evaluations are from rational reasoning, tangible beliefs, and utilitarian attitudes. For example, the following sentences express rational evaluations: "*The voice of this phone is clear*," "*This car is worth the price*," and "*I am happy with this car*."

Emotional evaluation: Such evaluations are from non-tangible and emotional responses to entities which go deep into people's state of mind. For example, the following sentences express emotional evaluations: "*I love iPhone*," "*I am so angry with their service people*" and "*This is the best car ever built*."

To make use of these two types of evaluations in practice, we can design 5 sentiment ratings, *emotional negative* (-2), *rational negative* (-1), *neutral* (0), *rational positive* (+1), and *emotional positive* (+2). In practice, neutral often means no opinion or sentiment expressed. Finally, we need to note that the concepts of emotion and opinion are clearly not equivalent. Rational opinions express no emotions, e.g., "*The voice of this phone is clear*", and many emotional sentences express no opinion/sentiment on anything, e.g., "*I am so surprised to see you here*".

More importantly, emotions may not have targets, but just people's internal feelings, e.g., "*I am so sad today*."

6. Recent Literature

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth,

User generated content on Twitter (produced at an monumental rate of 340 million tweets per day) provides a rich supply for gleaning people's emotions, which is necessary for deeper understanding of people's behaviors and actions. Extant studies on emotion identification lack comprehensive coverage of "emotional situations" as a result of they use comparatively little training datasets. To overcome this bottleneck, they have mechanically created an outsized emotion-labeled dataset (of concerning 2.5 million tweets) by harnessing emotion-related hashtags out there in the tweets. Authors have applied two totally different machine learning algorithms for feeling identification, to study the impactiveness of assorted feature combinations additionally because the effect of the dimensions of the training data on the feeling identification task. Their experiments demonstrate that a combination of unigrams, bigrams, sentiment/emotion bearing words, and parts-of-speech information is most effective for gleaning emotions. The highest accuracy (65.57%) is achieved with a training data containing concerning two million tweets.

Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio, Amir Hussain

Emulating the human brain is one of the core challenges of computational intelligence, which entails several key problems of artificial intelligence, including understanding human language, reasoning, and emotions. In this work, computational intelligence techniques are combined with common-sense computing and linguistics to analyze sentiment data flows, i.e., to automatically decrypt however humans categorical emotions and opinions via natural language. The increasing availability of social data is extraordinarily helpful for tasks such as branding, product positioning, corporate reputation management, and social media marketing. The elicitation of helpful info from this vast quantity of unstructured information, however, remains an open challenge.. In particular, they describe a novel paradigm for real-time concept-level sentiment analysis that blends machine intelligence, linguistics, and common-sense computing in order to enhance the accuracy of computationally expensive tasks like polarity detection from massive social information. The main novelty of their work consists in an formula that assigns discourse polarity to ideas in text and flows this polarity through the dependency arcs so as to assign a final polarity label to every sentence.

Otto K. M. Cheng, Raymond Lau , In the era of huge data, huge volumes of knowledge are generated from on-line social networks, sensor networks, mobile devices, and organizations' enterprise systems. This phenomenon provides organizations with unprecedented opportunities to faucet into big data to mine valuable business intelligence. However, traditional business analytics ways might not be able to address the flood of big data. The main contribution authors is that the illustration of the event of a completely unique big data stream analytics framework named BDSASA that leverages a probabilistic language model to investigate client the buyer the patron sentiments embedded in many countless on-line consumer reviews. In particular, an logical thinking model is embedded into the classical language modeling framework to enhance the prediction of shopper sentiments.

Bingwei Liu, Erik Blaschy, Yu Chenz, Dan Shen_ and Genshe Chen
A typical method to get valuable information is to extract the sentiment or opinion from a message. Machine learning technologies are wide used in sentiment classification attributable to their ability to "learn" from the training dataset to predict or support deciding with comparatively high accuracy. However, when the dataset is massive, some algorithms might not proportion well. In this paper, authors aim to value the measurability of Na¨ıve bayes classifier (NBC) in massive datasets. Instead of employing a standard library (e.g., Mahout), authors implemented NBC to come through fine-grain management of the analysis procedure. A Big knowledge analyzing system is additionally style for this study. The result is encouraging therein the accuracy of NBC is improved and approaches 82% once the dataset size will increase. They have demonstrated that NBC is ready to scale up to analyze the sentiment of millions movie reviews with increasing throughput.

Ahmad Ghazal1,Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen There is an incredible interest in big data by academe, industry and a massive user base. Several industrial and open supply suppliers unleashed a selection of product to support big data storage and process. As these products mature, there is a requirement to judge and compare the performance of those systems. In their work authors present BigBench, an end-to-end big data benchmark proposal. The underlying business model of BigBench is a product retailer. The proposal covers a data model and synthetic data generator that addresses the range, velocity and volume aspects of big data systems con-taining structured, semi-structured and unstructured data.

Seref SAGIROGLU and Duygu SINANC Big data is a term for {large} data sets having large, more varied and complicated structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of analysis into large amounts of {data of knowledge of information} to reveal hidden patterns and secret correlations named as big data analytics. These useful information's for companies or organizations with the facilitate of gaining richer and deeper insights and obtaining a plus over the competition. For this reason, big information implementations would like to be analyzed and executed as accurately as potential. **Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg** Sentiment analysis research has been increasing hugely in recent times due to the big selection of business and social applications. Sentiment analysis from unstructured natural language text has recently received considerable attention from the analysis community. This work propose a novel sentiment analysis model supported common-sense information extracted from construct internet based mostly ontology and context data. Concept internet based mostly ontology is used to see the domain specific ideas that successively created the domain specific vital options. Further, the polarities of the extracted concepts are determined exploitation the discourse polarity lexicon that we have a tendency to developed by considering the context data of a word. Finally, semantic orientations of domain specific options of the review document are aggregative based mostly on the importance of a feature with relation to the domain. The importance of the feature is determined by the depth of the feature within the ontology.

Marcos D. Assun, Rodrigo N. Calheirosb, Silvia Bianchic, Marco A. S. Nettoc, Rajkumar Buyyab
In their work authors discusses approaches and environments for carrying out analytics on Clouds for big Data applications. It revolves around four important areas of analytics and big data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualisation and user interaction; and (iv) business models. Through a detailed survey, we determine potential gaps in technology and offer recommendations for the analysis community on future directions on Cloud-supported massive information computing and analytics solutions.

Yang Yu, Xiao Wang,
The present project collected period tweets from U.S. soccer fans during five 2014 FIFA World Cup games (three games between the U.S. team and another opponent and two games between alternative teams) exploitation Twitter search API. They used sentiment analysis to examine U.S. soccer fans' emotional responses in their tweets, particularly, the emotional changes after goals (either own or the opponent's). Authors found that throughout the matches that the U.S. team played, fear and anger were the most common negative emotions and generally, increased once the opponent team scored and bated once the U.S. team scored. Anticipation and joy were also usually consistent with the goal results and therefore the associated circumstances throughout the games. Furthermore, they found that throughout the matches between alternative teams, U.S. tweets showed more joy and anticipation than negative emotions (e.g., anger and fear) and that the patterns in response to goal or loss were unclear. This project revealed that sports fans use Twitter for emotional functions and that the large data approach to analyze sports fans' sentiment

showed results usually according to the predictions of the disposition theory once the fanship was clear and showed good predictive validity.

7. Conclusion

This chapter defined the concept of opinion in the context of sentiment analysis, the main tasks of sentiment analysis, and the framework of opinion summarization. Along with them, two relevant and important concepts of subjectivity and emotion were also introduced, which are highly related to but not equivalent to opinion. Existing studies about them have mostly focused on their intersections with opinion (although not always). However, we should realize that all these concepts and their definitions are rather fuzzy and subjective. For example, there is still not a set of emotions that all researchers agree. Opinion itself is a broad concept too. Sentiment analysis mainly deals with the evaluation type of opinions or opinions which imply positive or negative sentiments.

References:

[1]. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth, “Harnessing Twitter ‘Big Data’ for Automatic Emotion Identification”, http://blog.twitter.com/2012/03/twitter-turns-six.html

[2]. Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio, Amir Hussain, ARTICLE in IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE, October, 2015

[3]. Otto K. M. Cheng, Raymond Lau, (2015), “Big Data Stream Analytics for Near Real-Time Sentiment Analysis”, **Journal of Computer and Communications, 3, 189-195**

a. Published Online in SciRes. http://www.scirp.org/journal/jcc

b. http://dx.doi.org/10.4236/jcc.2015.35024

[4]. Bingwei Liu, Erik Blaschy, Yu Chenz, Dan Shen_ and Genshe Chen, (2013), “Scalable Sentiment Classification for Big Data Analysis Using Naıve Bayes Classifier”, **Big Data, IEEE International Conference, pp. 99-104.**

[5]. Ahmad Ghazal1,Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen, (2013), “BigBench: Towards an Industry Standard Benchmark for Big Data Analytics”, *SIGMOD’13*, Copyright 2013 ACM 978-1-4503-2037-5/13/06

[6]. Seref SAGIROGLU and Duygu SINANC, (2013), “Big Data: A Review”, 978-1-4673-6404-1/13/©2013 IEEE

[7]. Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg, (2015), “Sentiment Analysis Using Common-Sense and Context Information**”,** Computational Intelligence and Neuroscience, Volume 2015, Article ID 715730, 9 pages

a. http://dx.doi.org/10.1155/2015/715730

[8]. Marcos D. Assun, Rodrigo N. Calheirosb, Silvia Bianchic, Marco A. S. Nettoc, Rajkumar Buyyab, (2014), “Big Data Computing and Clouds:Trends and Future Directions”, arXiv:1312.4722v2 [cs.DC]

[9]. Utkarsh Srivastavaa, Santosh Gopalkrishnanb, (2015), “Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks“,2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), Elsevier Procedia Computer Science 50 (2015) 643 – 652

[10]. Amir Gandomi, Murtaza HaiderTed, (2015), “Beyond the hype: Big data concepts, methods, and analytics”, Elsevier International Journal of Information Management 35 (2015) 137–144

[11]. Yang Yu, Xiao Wang, (20158), “World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans’ tweets”, Elsevier Computers in Human Behavior 48 (2015) 392–400

[12]. Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter, 2006. 8(1): p. 41-48.

[13]. Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target dependent twitter sentiment classification. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011.

[14]. Jindal, Nitin and Bing Liu. Identifying comparative sentences in text documents. in Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006). 2006a.

[15]. Jindal, Nitin and Bing Liu. Mining comparative sentences and relations. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006). 2006b.