DOI: 10.37943/AITU.2021.57.68.005

A. Mukasheva
Master's student, Information Systems Department
asselmukasheva17@gmail.com, orcid.org/0000-0002-3695-0978
L.N. Gumilyov Eurasian National University, Kazakhstan

TASKS AND METHODS OF TEXT SENTIMENT ANALYSIS

Abstract: The purpose of this article is to study one of the methods of social networks analysis – text sentiment analysis. Today, social media has become a big data base that social network analysis is used for various purposes – from setting up targeted advertising for a cosmetics store to preventing riots at the state level. There are various methods for analyzing social networks such as graph method, text sentiment analysis, audio, and video object analysis. Among them, sentiment analysis is widely used for political, social, consumer research, and also for cybersecurity. Since the analysis of the sentiment of the text involves the analysis of the emotional opinions expressed in the text, the first step is to define the term opinion. An opinion can be simple, that is, a positive, negative or neutral emotion towards a particular object or its aspect. Comparison is also an opinion, but devoid of emotional connotation. To work with simple opinions, the first task of text sentiment analysis is to classify the text. There are three levels of classifications: classification at the text level, at the level of a sentence, and at the aspect level of the object. After classifying the text at the desired level, the next task is to extract structured data from unstructured information. The problem can be solved using the five-tuple method. One of the important elements of a tuple is the aspect in which an opinion is usually expressed. Next, aspect-based sentiment analysis is applied, which involves identifying aspects of the desired object and assessing the polarity of mood for each aspect. This task is divided into two sub-tasks such as aspect extraction and aspect classification. Sentiment analysis has limitations such as the definition of sarcasm and difficulty of working with abbreviated words.

Keywords: sentiment analysis, opinion, aspect, unstructured text, structured data, classification.

Introduction

Social networks today provide researchers from different fields with the opportunity to analyze different users in detail. Now there are different social networks – Instagram, Facebook, Twitter, WeChat, etc. Among them, the most analyzed social network is Twitter. For example, in the United States, there is a project called Pulse of the Nation [1], which deals with determining the mood of Americans who actively use Twitter during the day. The USA also has a SportSense project [2], which measures the level of excitement of football fans from their tweets in order to track important moments of the game in real time. Social media analysis has also been used to identify potentially dangerous personalities after the 9/11 attacks in the United States. A group of scientists, analyzed the emotional coloring of tweets (messages on Twitter) and on this basis created the vocabulary of terrorists, declared to be dangerous user accounts [3].

The abovementioned projects use the method of *text sentiment analysis* (sentiment analysis) to solve the problems, which is one of the methods for analyzing social networks. This method solves such problems as identifying emotionally colored text and its analysis. That is, text sentiment analysis is a must for big data and computational linguistics. The article discusses the basic terms of problems of text sentiment analysis, classification, methods of converting unstructured into structured data, as well as a sentiment analysis method based on aspects of an object.

Main body

According to the definition given in [4]: "the analysis of the sentiment of the text is the task of the automatic analysis of opinions and emotionally colored vocabulary expressed in the text."

According to the definitions, textual information on the Internet is divided into two classes: facts and opinions [4]. Defining opinion is one of the most important points,- because it affects the further construction of the algorithm for analyzing the sentiment of the text.

There is a simple opinion and comparison.

Examples of a simple opinion might be "I was pleasantly surprised by the quality of the furniture assembly" or "After the course of therapy, my health improved." In the first case, the author spoke directly about one object, and in the second case, the author's statement is implicit. But in both the first and second examples, one can notice that the texts have a positive emotional coloring [5].

A simple opinion has five elements:

- entity (object, topic) – object,
- feature (aspect, facet) – aspect,
- holder (opinion source) – author
- time – the moment in time when the opinion was expressed,
- sentiment value: positive, negative and neutral, that is, without emotional coloring, types of emotions [5].

Examples of comparisons might be "The new Samsung is more expensive than the new IPhone", "Both the new Samsung and the new IPhone have a high-quality camera," or "Xiaomi has the best home products". The first example shows a comparison of aspects of different objects, the second example equates aspects of different objects, and the third example shows the superiority of one object. That is, the comparison is divided into three categories as:

- Uneven gradation
- Equivalent
- Excellent degree

The comparison also has five elements:

- entity 1, entity 2 – objects to be compared,
- aspect – the aspect of the comparison,
- po – preferred objects,
- holder – author,
- time – the moment in time of expressing the opinion.

It should be noted that in the comparisons there is no emotional coloring of the text [6].

After defining what an opinion is, consider at what levels sentiment analysis can classify a simple opinion. There are three levels of classification: at the document level, at the proposal level or at the aspect level [7].

Classification in a level of document

The process of classification in a level of the document implies extracting sentiment from the whole opinion, then overall sentiment of the opinion-holder leads to the classification of overall opinion. The aim of this process is to assess if opinion can be classified into one of the categories – positive, negative, or neutral. For example, let's consider the opinion "A few days ago I bought a new IPhone. This is a very nice phone, albeit a bit big. The touchscreen is awesome. The voice quality is also good. I really like!" Clearly, this opinion will be classified as *positive*. The process of classification will lead to better results when an opinion-holder writes simple opinion, meaning one author – one object.

Classification by proposal level

This process usually consists of two stages: the first stage – subjective classification and the second stage – sentimental classification. In the first stage it is determined if the proposal is objective or subjective, in the second stage, obviously, it is classification of sentiment into positive or negative.

Classification of the proposal into objective or subjective is based on the information given in a sentence. Therefore, the proposal is classified as objective if it contains more factual information. On the other hand, the proposal is classified as subjective if it contains personal feelings and beliefs. Usually, it is harder to identify subjective sentences and so one of the methods is using naive Bayesian classification. This is a preparation step that helps to sort out sentences without opinion, determine the tone of the text with an opinion that contains the subject and its aspect. Generally, proposals with subjective classification can consist of several opinions, can contain both objective and subjective information. A subjective proposal can contain several opinions and subjective and factual information. For example, "Tourism is still doing well in this pandemic period." One of the drawbacks of sentiment classifications in a level of document and by proposal level is that it is hard to determine what people prefer, why they write opinion, and what they dislike.

Converting unstructured text to structured data

The next task is to structure the information. That is, in computational linguistics, texts written in human language are considered unstructured. Sentiment analysis is used to convert text into structured information and analyze it for further use.

There are various techniques for getting structured data from unstructured information. For example, a tuple of five elements (e_j, a_{jk}, so_{ijkl}, h_i, t_i) method can be used, where:

e_j – target (object)

a_{jk} – aspect / characteristic of the object e_j

so_{ijkl} – is the emotional coloring of the author's opinion

so_{ijkl} can be positive, negative or neutral, or if there are no other indicators

h_i – opinion author

t_i – time when the opinion was published [8].

Target e_j is not just an object, it might be a person, for example, celebrities, or a topic, for example covid19, or hierarchy of components and so on. For example, review from ABC on 07/08/2021 – "I bought an IPhone 11 smartphone. As they say, this is a great phone. The screen resolution is great and the storage capacity is big." This opinion can be expressed as follows:

Table 1. Elements of five-element tuple

e_j	a_{jk}	so_{ijkl}	h_i	t_i
IPhone 11	screen	+	ABC	08.07.2021
IPhone 11	storage	+	ABC	08.07.2021

(IPhone 11, screen, +, ABC, 08.07.2021)
(IPhone 11, storage, +, ABC, 08.07.2021)

This example shows how to get structured data from unstructured information. The method used helped us to get a structured summary of sentences. As soon as structured data is extracted, the next step can be sentiment analysis of formatted data.

Sentiment analysis in a level of aspect

The process of sentiment analysis in a level of aspect is based on recognizing aspects of the desired object, then assessing the tone of the emotion regarding each identified aspect. This process consists of two sub-tasks: extraction and classification.

Aspect extraction refers to the recognition of aspects of an object, and basically, it is a task of extracting information. The second subtask – classification – identifies positive, negative or neutral opinions on different aspects.

While vocabulary-based approaches use special lists of emotional phrases associated with an aspect as the main resource, the key to solving the problem is to recognize the limits and boundaries of each specific expression of mood that relates to the aspect of the object.

Extracting Aspects

To define all terms of that are in a sentence, every common phrase (e.g. food) must be found and filtered according to the rule "comes immediately after emotional / sentimental / word" (e.g. delicious food). Then, based on this, a set of common phrases are be built [9]. An alternative way is to build initially defined aspects and search for them throughout the proposals. For a delivery service, these could be the following aspects: cost, time, service, quality. For example, "The delivery time was really short, but it cost me a lot". Aspects: time, cost.

Sentiment classification of aspects

Texts that contain words in them can express a wide range of feelings that vary from positive to negative, and these feelings may be strongly expressed or weak. Determining the polarity of words and identifying to which categories they belong is a crucial part of sentiment analysis. One way to tackle this task is to use existing lexicons that are available and already have moods classified into categories like positive or negative [10].

Available lexicons have a number of limitations because they may be missing words or existing words may not be relevant to the theme of discussion. If that happens, a new lexicon can be created. This can be achieved through semi-controlled vocabulary, that is based on input of a small amount of information (for example, some manually downloaded instances) and, based on the information available, creates a complete vocabulary. In another approach, bootstrapping, a machine is used to accurately classify tone of the sentence as subjective and objective. A batch of templates is then extracted from these sentences. The process of extracting more subjective and objective sentences is then repeated over and over again to reach the desired vocabulary.

DOI: 10.37943/AITU.2021.57.68.005
© A. Mukasheva

Finding the tonality of an aspect

This task is considered through an example. For example, one customer posted a review of a restaurant: “The delivery time was really short, but it cost me a lot.” The diagram below shows the components of a system that is used to find aspect sentiment.

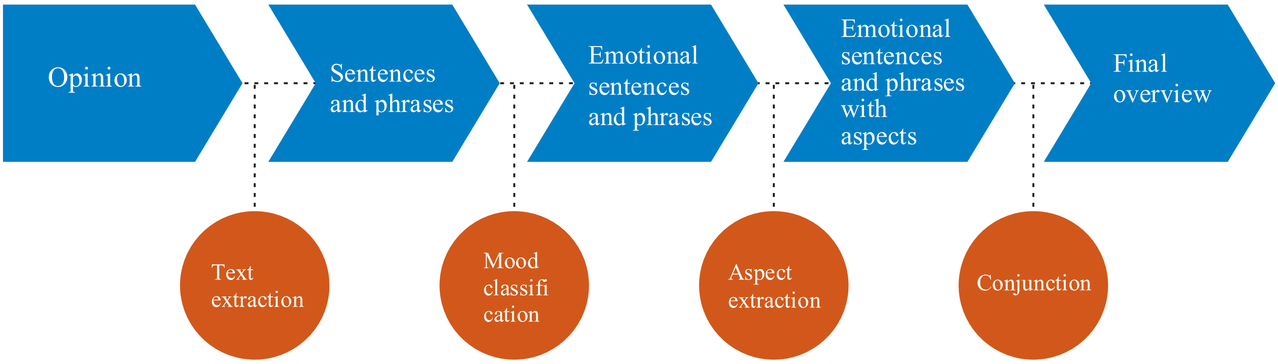

Fig. 1. System components for finding the sentiment of an aspect

Why use sentiment analysis?

A significant portion of the world’s data is presented in text form, for example: emails, posts on social media, articles, documents. Text data is inherently unstructured, which makes it very difficult to process, but at the same time, text data contains a lot of useful knowledge. Therefore, interest in automatic text analysis systems is steadily increasing. Opinion analysis systems, which belong to this class of systems, allow companies to automatically extract useful knowledge from text data, which, in turn, saves hours of manual labor and automates many business processes [11].

The advantages of sentiment analysis systems include the following:

1. Scalability. It is impossible to imagine working with millions of social media posts manually. Sentiment analysis is a time-efficient tool that gives us an opportunity to process large amounts of data with a minimum cost. The increase in data processing leads to a slight increase in cost caused by the purchase of additional disk space and computing power.

2. Analysis in real-time. Sentiment analysis can be used to identify critical information that provides real-time information awareness in specific situations. A sentiment analysis system can identify early PR crises and unsatisfied customers.

3. Coordination of evaluation criteria. When people are involved in assessing the sentiment of statements, even one person, depending on different factors (mood, attitude to the topic, ...), can give a different assessment of the same messages. The problem is compounded when several people are involved in the task of assessing user opinions. In this case, it is extremely difficult to reconcile the ratings of two different evaluators: one person can evaluate the opinion as positive, the other as neutral, etc.

Sentiment analysis application

Increasing interest towards sentiment analysis is conditioned by accessibility of various information from the web, which contains emotional opinion. Sentiment analysis has developed to that extent that it is widely used in different fields to learn about customers, and analyze their preferences, what they like and dislike. Some of these applications are described below.

Business applications

Analysis of emotionally charged texts has been exploited by many companies seeking to take advantage of "market sentiment". Sentiment analysis can be used to research products, track brands, change marketing strategies, and achieve financial innovation. The main activities supported by sentiment analysis are:

- Automatic tracking of brands, products, and services ratings from combined user reviews and review sites;
- Analysis of trends of buyers, competitors, and market trends;
- Assess the company's response to events taking place around. When something new is launched, sentiment analysis can provide immediate information on product acceptance.
- Can evaluate whether a brand image is liked or disliked.
- Monitoring of critical issues to prevent harmful virus effects.

The main problems identified by researchers in these applications are:
- identify aspects of the product;
- link feedback with product aspects;
- identify fake reviews and process non-compliant reviews.

Applications in policy

Sentiment analysis allows you to follow the problems and subjective opinions of bloggers in specific blogs with politics thematic. It can be a tool for political organizations that is used to assess what issues are close to voters. The authors determined whether the proposed legislation was supported in the transcript of the debates in the US Congress. In order to increase the importance of the information provided to voters, it is possible to determine the weight of public figures, the reasons why they choose or do not choose.

Recommender system

These systems can help extract user ratings from text. Such platforms as Netflix, Ivy or Kinopoisk classify film reviews as "you might like" and "might not like".

Expert finding

It is a method of tracking literary influence. For example, if each group member gain knowledge from the online society using a blog and sentiment analysis of the comments obtained, then combined navigation power determines the blog account that allows the blogs to be ranked and determines whether the expert is the best in the ranking system.

Summarization

Summarizing emotionally charged text applies when a scale of online views of an object is huge. This creates difficulty for the consumers, because they can be lost among thousands of reviews losing the possibility to make reasonable decision. This also creates difficulty for the product as well, because producer may also get lost among opinions and it gets impossible to follow them. For example, research on specific products can be conducted:

1) the detected features of the product were explained;
2) opinion sentences were defined for each feature;
3) a summary was prepared using the information found.

Completion of one or more documents is an application that can improve sentiment analysis.

Government intelligence

An increase in antagonistic or hostile relations can be observed with this area of application suggested by monitoring sources.

Sentiment analysis problems

Despite its widespread use, sentiment analysis has a number of limitations. For example, human judgment is still the most accurate instrument for communicating emotions. Let's have a closely look at them.

Keyword selection

A pool of keywords is usually used to classify opinions. In analysis of emotionally charged texts, text is classified according to two very different classes (positive and negative). But finding the right set of keywords is not an easy task. Moods can often be expressed with sensitivity, which can make it difficult to consider an expression in a sentence or document separately. For example, "If you think this is your favorite perfume, please use it at home and close the windows" does not contain any negative words.

Sentiment is domain specific

Opinion depends on the subject area and the meaning of the words varies based on the content used. For example, a short statement "go read a book" is positive for a review of a book, but negative for a review of a film, that is, it has the opposite effect.

Various opinions in one proposal

There might be many opinions in a sentence, along with objective and subjective parts. Such parts should be separated for further use. For example, this opinion has both positive and negative tones "The new Kindle's battery life is long, but the is very small".

Negation handling

Repudiation in sentiment analysis is difficult to manage. For instance, texts "I love Americano" and "I don't love Americano" are distinct only in a single feature, so they should be classified as different and opposite. The main difficulty is in the use of other languages, negative expressions, product features or attributes, sentence or document complexity, hidden product features, and so on. is to prepare a summary of opinions related to.

Sarcasm

Irony and sarcasm are common in political and online discussions. These words vary from language to language, making them difficult to identify. Very little research has been done on this topic.

Implicit opinion

Comments can be classified as explicit and implicit: "We had a great time" is clear, and "The battery lasted three hours" is explicit. Existing sentiment analysis models cannot detect a secretly expressed negative tone of the opinion.

Comparative sentences

There is limited research on the classification of comparative sentences as opinionated or not. At the same time, the order of the words in the comparative proposals clearly shows difference in determining the orientation of the opinion. For example, the opinion "IPhone is better than Samsung" means the exact opposite of the sentence "Samsung is better than IPhone."

Multilingual sentiment analysis

Much of the work on sentiment analysis has concentrated on English-language data, mainly due to the availability of resources such as dictionaries and hand-marked corpses. Because the vast majority of Internet users speak English, there is a need to create materials and research in other languages. Some researchers already suggested several methods for using English-language resources and tools using interlingual projects [12].

Noisy texts

Spelling and grammatical errors, omitted or problematic punctuation, and jargon remain a problem in most sentiment analysis systems.

Opinion spam

This problem refers to falsified and misleading comments which deliberately seek to misinform users by giving false positive or false negative feedback to a specific object for the sake of advocating negative reputation of other objects. There are tools that determine spam based on those reviews, allowing the reviewer to evaluate the usefulness of the review for each review.

Conclusion

The growth in the volume of unstructured textual data stimulates interest in the problems of natural language analysis and in sentiment analysis in particular. The availability of open libraries of machine learning allows the use of modern algorithms to solve the problems of this class.

The use of sentiment analysis opens up new business opportunities. Automatic sentiment detection will allow you to quickly and cheaply conduct research on social media. The proposed approach can be used to conduct marketing, sociological and political research. It also allows monitoring audience loyalty to a specific topic or brand, which enables management to make the necessary decisions in a timely manner.

References

1. Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010, May). Measuring user influence in twitter: The million-follower fallacy. In *Proceedings of the international AAAI conference on web and social media* (Vol. 4, No. 1).
2. SportSense [HTML](ec2.compute1.amazonaws.com/sportsense/)
3. Du, H., & Yang, S. J. (2011, March). Discovering collaborative cyber attack patterns using social network analysis. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 129-136). Springer, Berlin, Heidelberg.
4. Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2* (1-2), 1-135.
5. Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627-666.
6. Jindal, N., & Liu, B. (2006, July). Mining comparative sentences and relations. In *Aaai* (Vol. 22, No. 13311336, p. 9).
7. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews. (http://www.ryanmcd.com/papers/local_service_summ.pdf)
8. Andrey, A.K. (2015). Text search. Working with unstructured data. *Mathematics and information technology in the oil and gas complex*, (2), 115-126.
9. Sentiment Analysis by Professor Dan Jurafsky (https://web.standford.edu/class/cs124/lec/sentiment.pdf)
10. Semina, T.A. (2020). Sentiment analysis of the text: modern approaches and existing problems. *Social and Human Sciences. Domestic and foreign literature, Linguistics: Abstract Journal*, 6(4), 47-64.
11. Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia computer science*, *130*, 660-666.
12. Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.