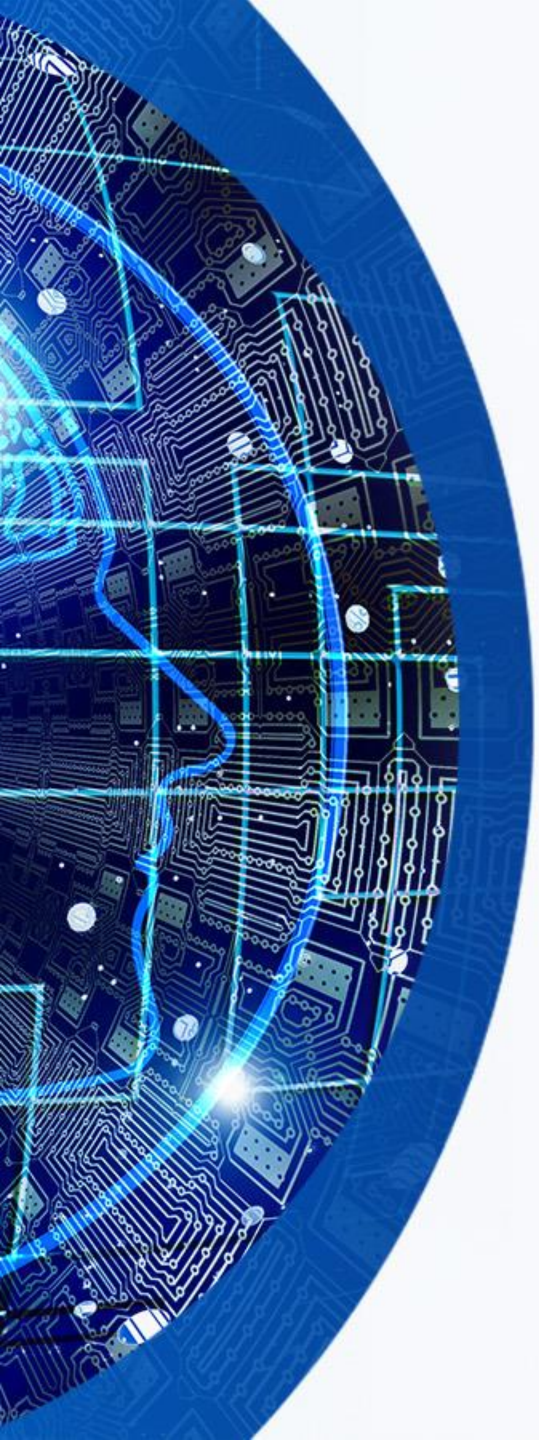
Natural Language Processing Applications

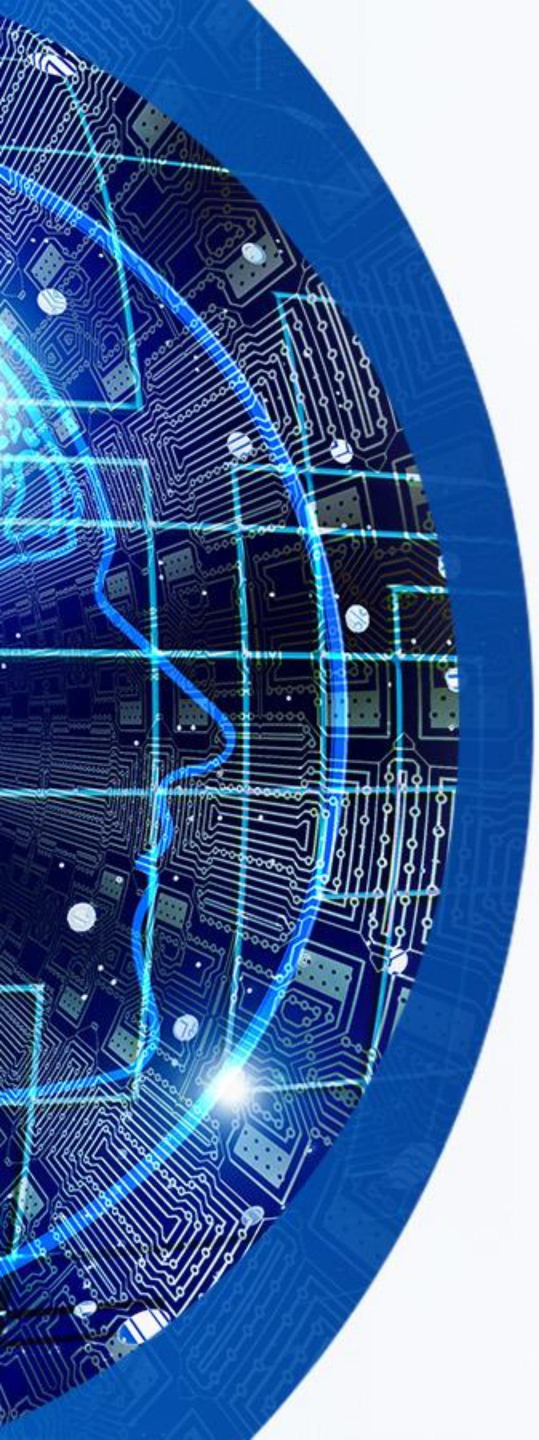
Information Extraction

Sentiment Analysis & Subjectivity



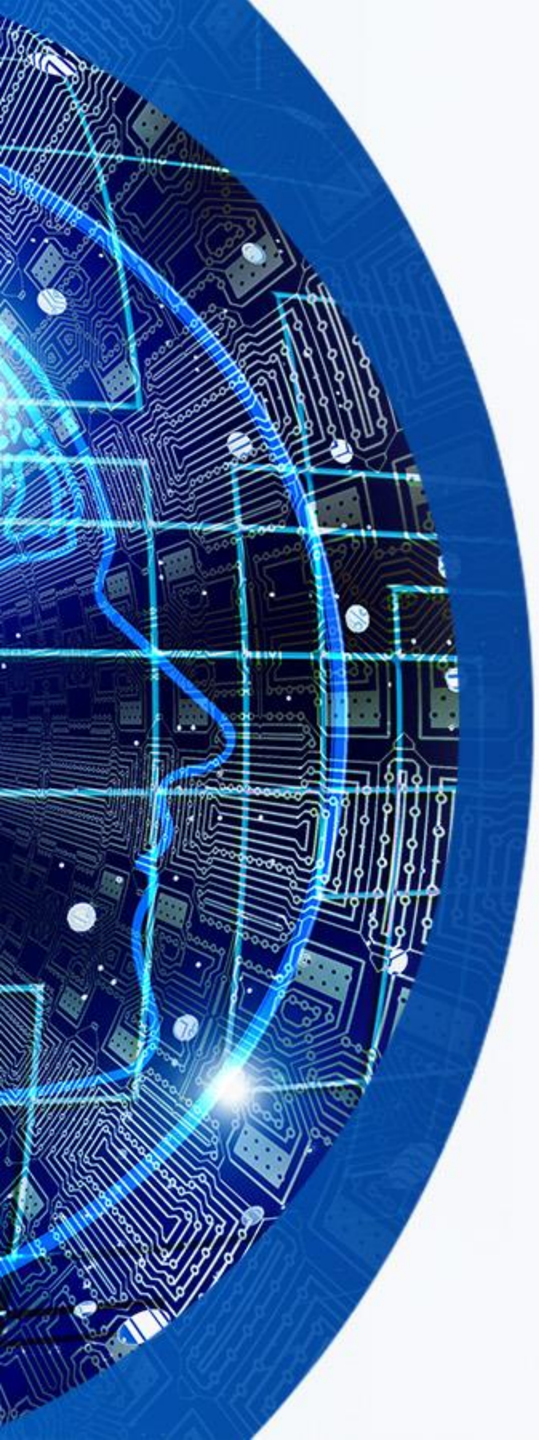
Outlines

- Information Extraction (IE)
- Sentiment Analysis (SA) and Subjectivity



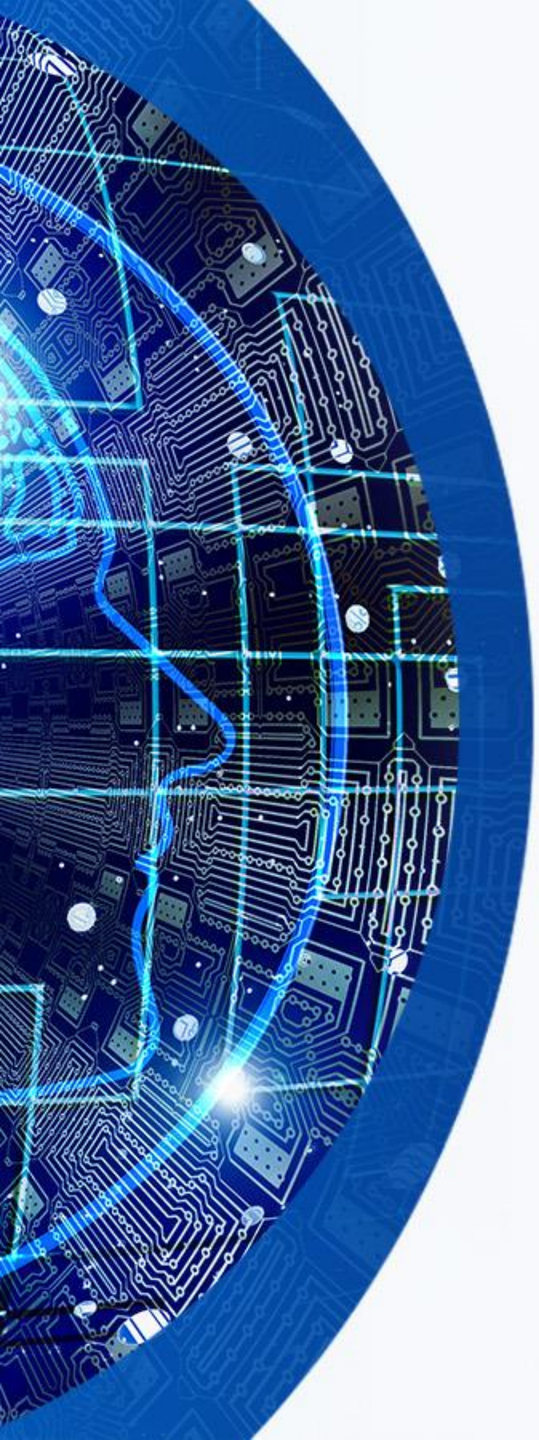
Information Extraction (IE)

- Information extraction (IE) is the process of **scanning** text for information **relevant** to some **interest**, including extracting entities, relations, and, most challenging, events—or who did what to whom, when, and where.
- It requires **deeper analysis** than keyword searches, but its aims fall short of the very hard and long-term problem of **text understanding**, where we seek to **capture** all the information in a text, along with the **speaker's or writer's intention.**
- IE represents a midpoint on this spectrum, where the aim is to capture structured information without sacrificing feasibility.



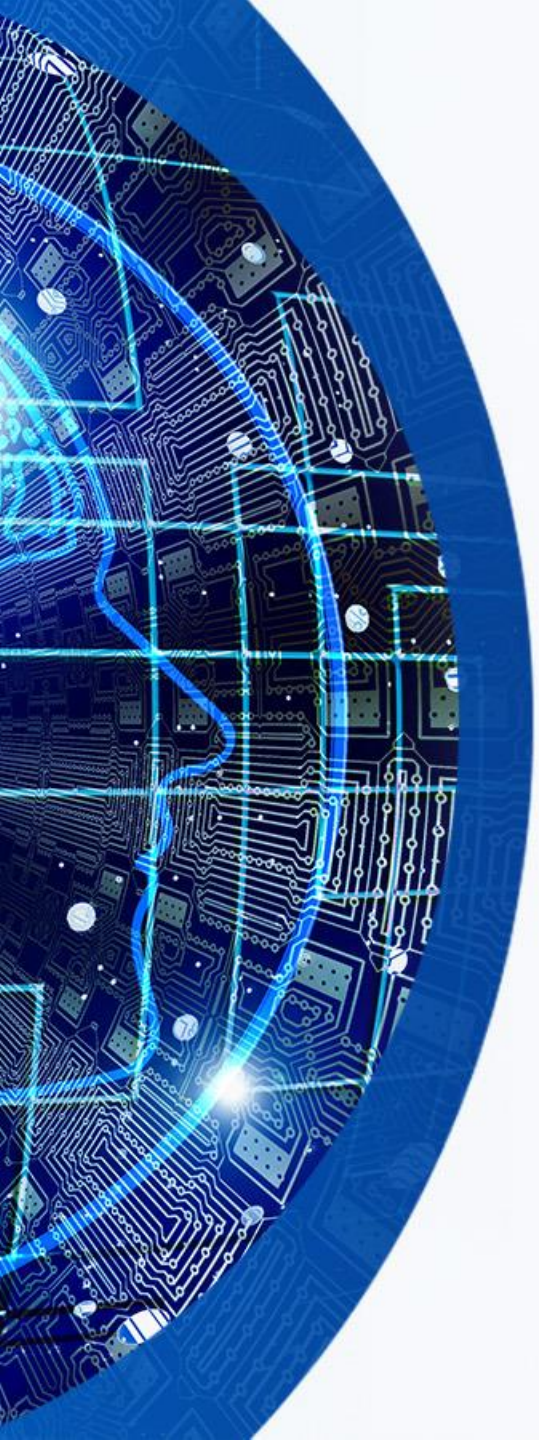
Information Extraction (IE)

- IE typically focuses on surface linguistic **phenomena** that do not require deep inference, and it focuses on the **phenomena** that are most frequent in texts.
- IE technology arose in response to the need **for efficient processing** of texts in specialized domains.
- IE technology focuses on only the **relevant parts** of the text and ignores the rest.



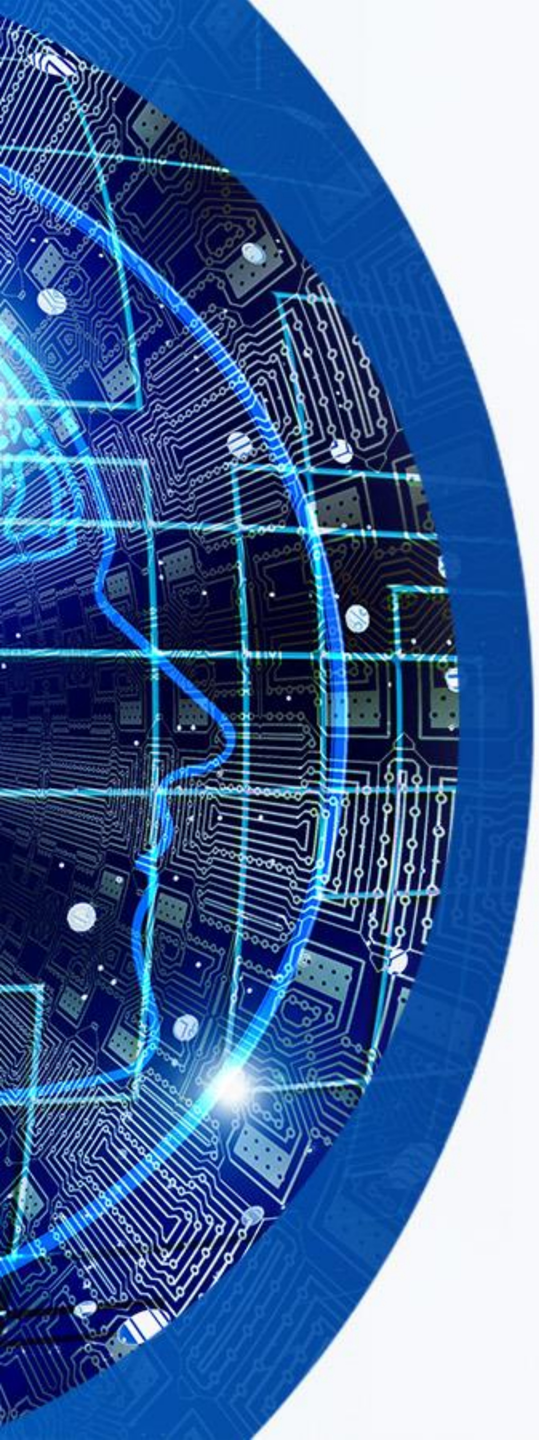
Information Extraction (IE)

- Typical applications of IE systems are :
  - **gleaning business**,
  - government,
  - or military intelligence from a large number of sources;
  - in searches of the **World Wide Web** for more specific information than keywords can discriminate;
  - for scientific literature searches;
  - in building databases from large textual corpora;
  - and in the curation of biomedical articles.



Information Extraction (IE)

- **Named entity recognition** (NER) is one of the most common uses of IE technology.
- NER is especially important in biomedical applications, where terminology is a formidable problem.
- NER systems identify different types of proper names,
  - such as **person and company names,**
  - and sometimes special types of entities, such as **dates and times**, that can be easily identified using surface level textual patterns.



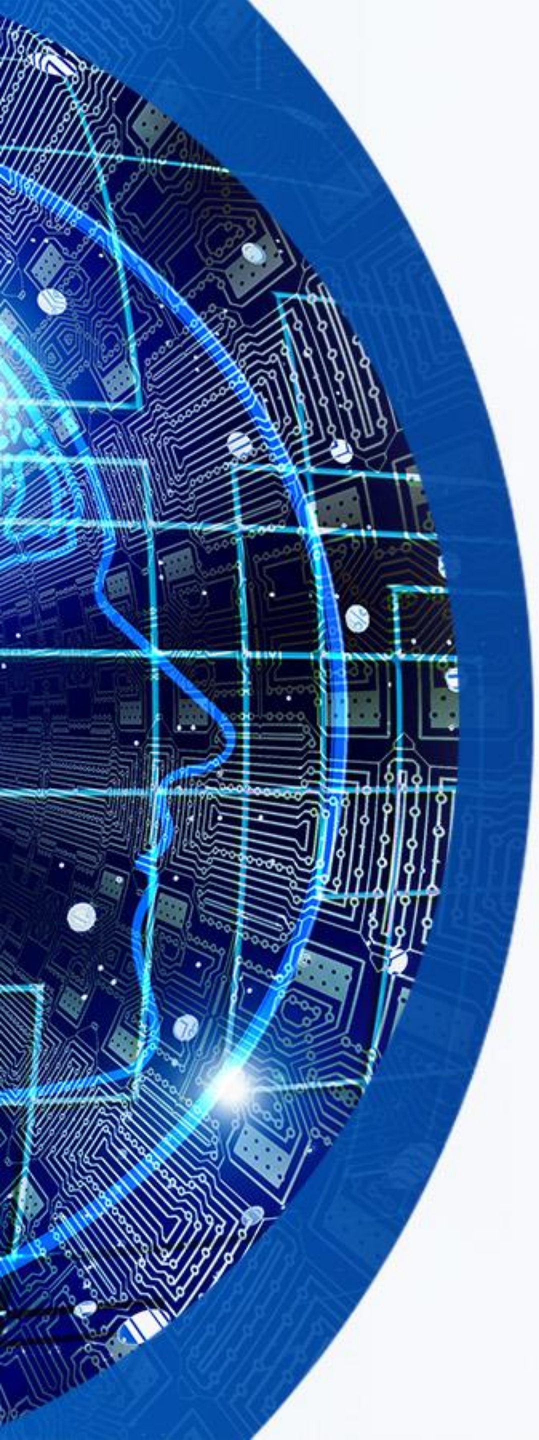
Information Extraction (IE)

- For example, in each of the sentences

  “Microsoft acquired Powerset.”

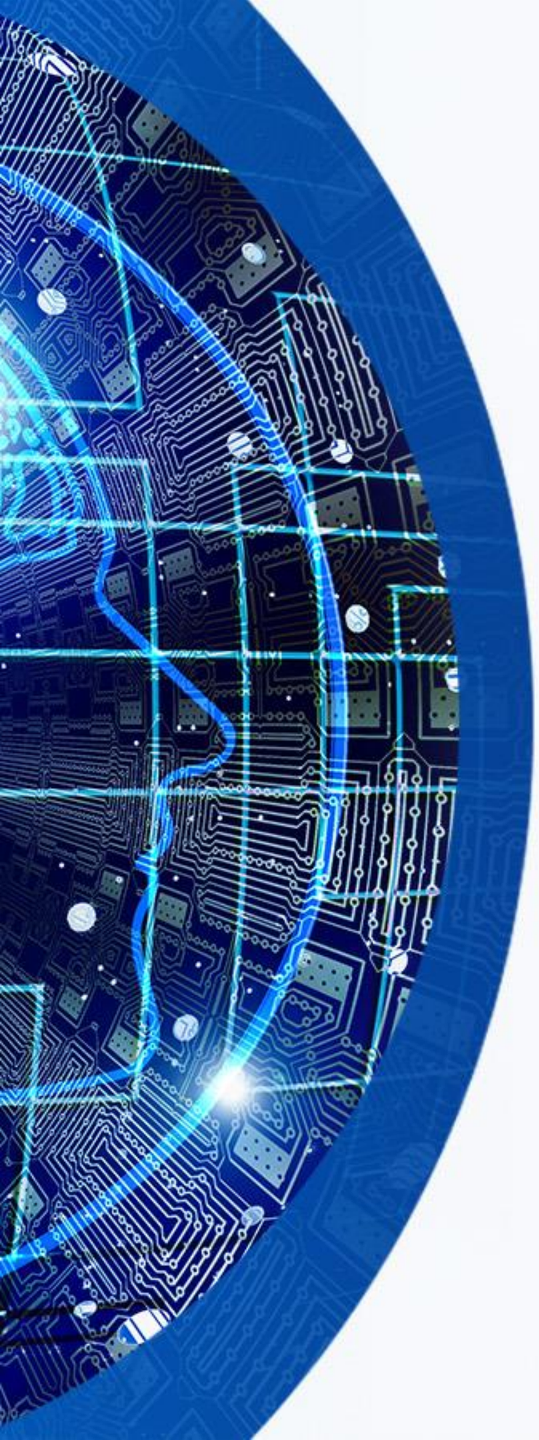
  “Powerset was acquired by Microsoft.”

- we would like to recognize not only that Microsoft and Powerset are company names, but also that an acquisition event took place, that the **acquiring company** was Microsoft, and the **acquired company** was Powerset.



Information Extraction (IE)

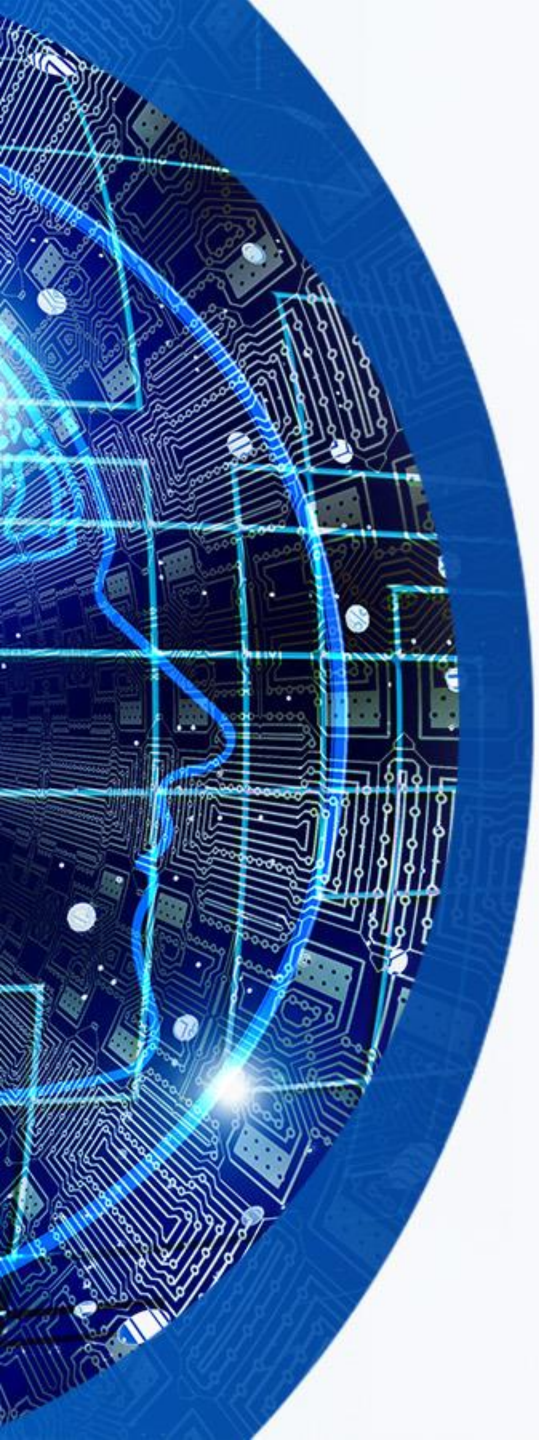
- Much of the technology in IE was **developed** in response to a series of evaluations and associated conferences called the **Message Understanding Conference** (MUC).
- Except for the earliest MUCs, these evaluations were based on a corpus of domain-specific texts, such as news articles on joint ventures.
- IE research has since been stimulated by the Automatic Content Extraction (ACE) evaluations.
- The ACE evaluations have focused on
  - identifying named entities,
  - extracting isolated relations,
  - and coreference resolution.



Information Extraction (IE)

- In a typical IE system, there are five levels of processing:

  1. **Complex Words**: This includes the recognition of multiwords and proper name entities, such as people, companies, and countries.
  2. **Basic Phrases**: Sentences are segmented into noun groups, verb groups, and particles.
  3. **Complex Phrases**: Complex noun groups and complex verb groups are identified.

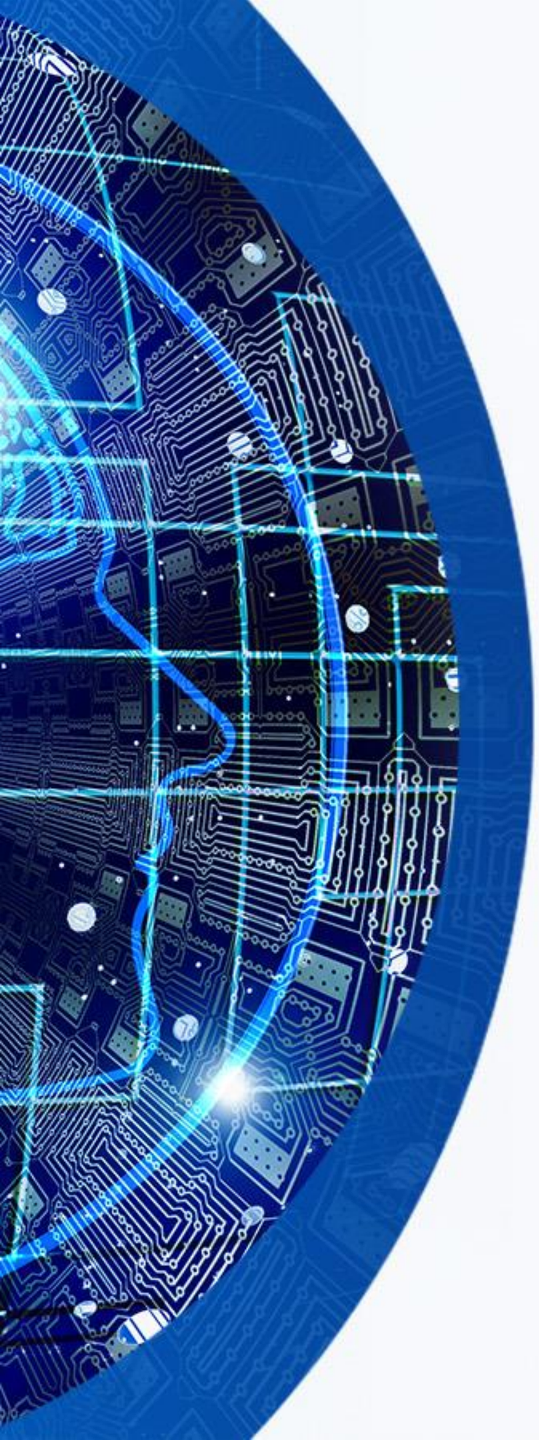


Information Extraction (IE)

4. **Domain Events**: The sequence of phrases produced at Level 3 is scanned for patterns of interest to the application, and when they are found, semantic structures are built that encode the information about entities and events contained in the pattern.

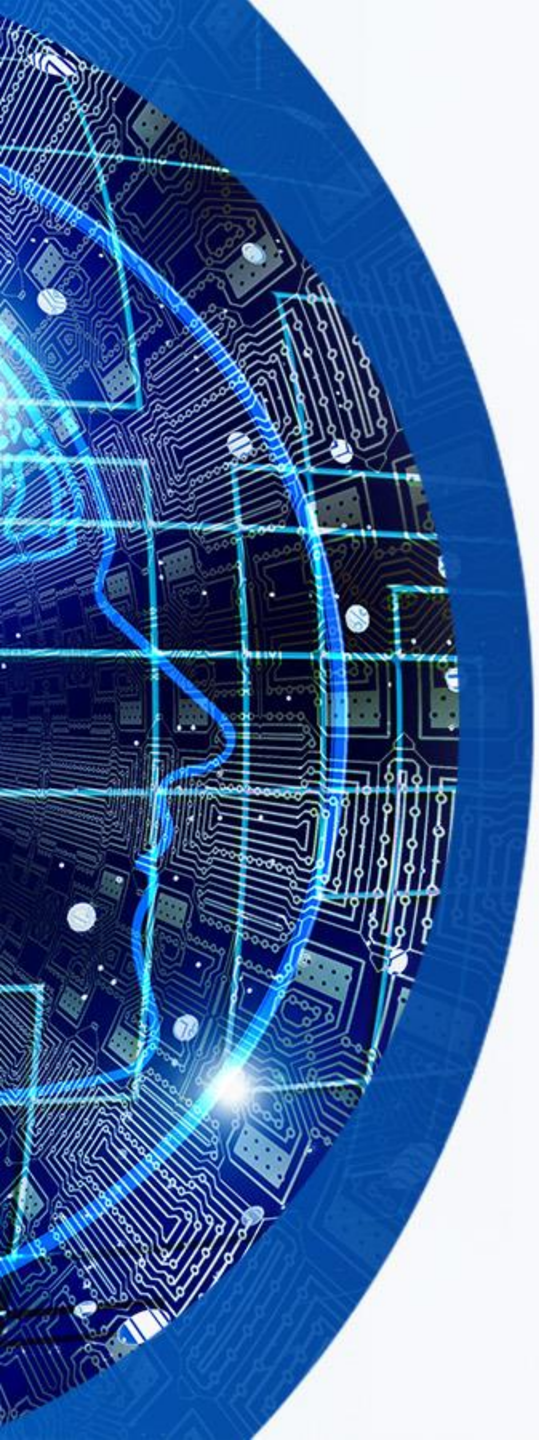
5. **Merging Structures**: Semantic structures from different parts of the text are merged if they provide information about the same entity or event.

This process is sometimes called template generation, and is a complex process not done by a finite-state transducer.



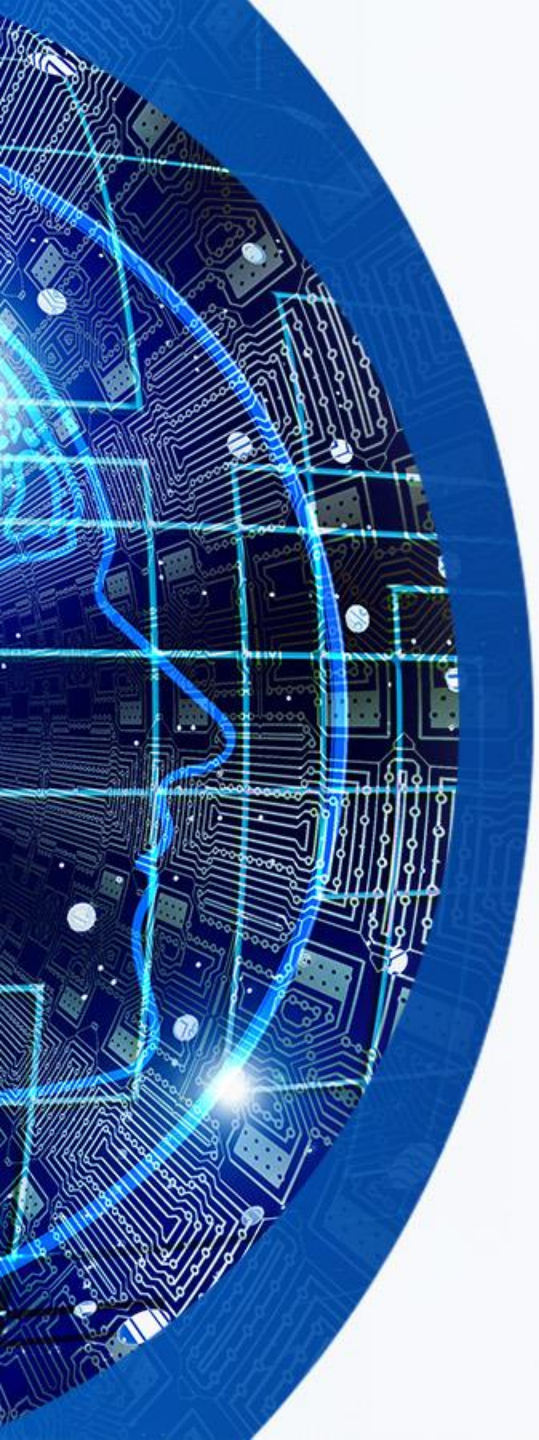
Sentiment Analysis (SA) and Subjectivity

- Textual information in the world can be broadly categorized into two main types:
    - ❑ **facts** and
    - ❑ **opinions.**
- Facts are **objective** expressions about entities, events, and their properties.
- Opinions are usually **subjective** expressions that describe people’s sentiments, appraisals, or feelings toward entities, events, and their properties.
- Much of the existing research on textual information processing has been focused **on the mining and retrieval of factual information**, e.g., information retrieval (IR), Web search, text classification, text clustering, and many other text mining and natural language processing tasks.



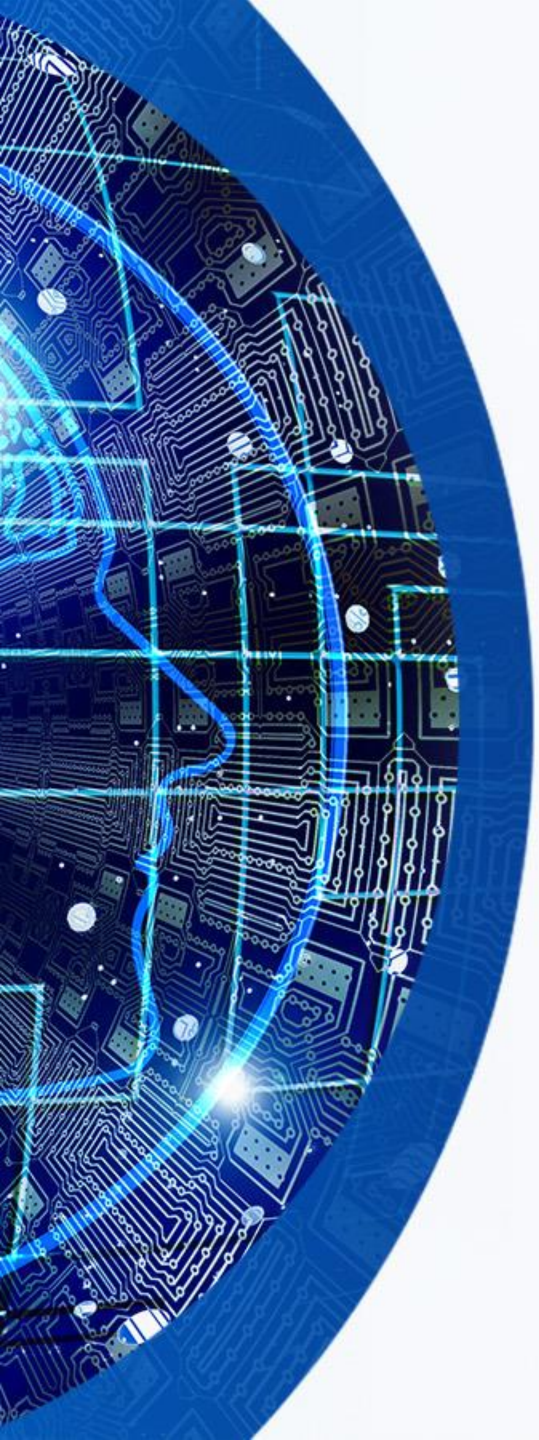
Sentiment Analysis (SA) and Subjectivity

- Little work had been done on the processing of **opinions** until only recently.
- Yet, opinions are so important that whenever we need to make a decision we want to hear others' opinions.
- This is not only true for **individuals** but also true for **organizations**.
- One of the main reasons for the lack of study on opinions is the fact that there was **little opinionated text** available before the World Wide Web.



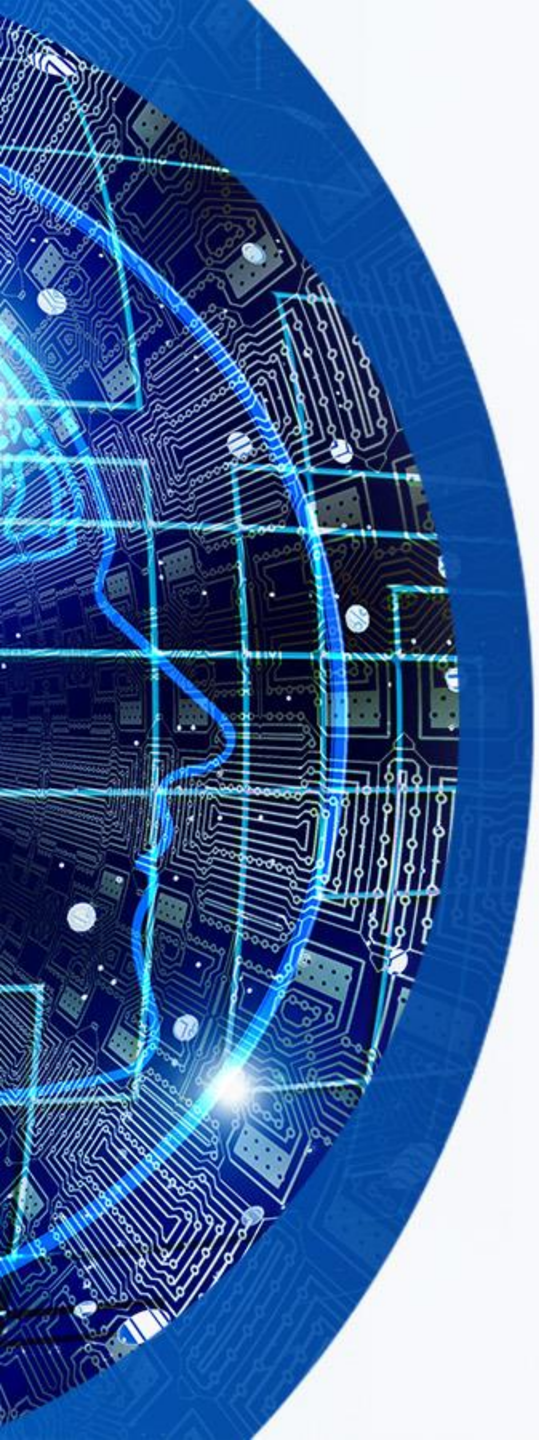
Sentiment Analysis (SA) and Subjectivity

- Before the Web, when an individual needed to make a decision, he or she typically asked for **opinions from friends and families.**
- When an organization wanted to find the opinions or sentiments of the general public about its products and services, **it conducted opinion polls, surveys, and focus groups.**
- However, with the Web, especially with the explosive growth of the **user-generated content** on the Web in the past few years, the world has been transformed.
- **The Web has dramatically changed the way that people express their views and opinions**.



Sentiment Analysis (SA) and Subjectivity

- They can now post **reviews** of products at merchant sites and express their views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the **user-generated content**.
- This online word-of-mouth behavior represents new and measurable sources of information with many practical applications.
- Now if one wants to buy a product, he or she is no longer limited to asking his or her friends and families because there are many **product reviews on the Web that give opinions of existing users of the product.**

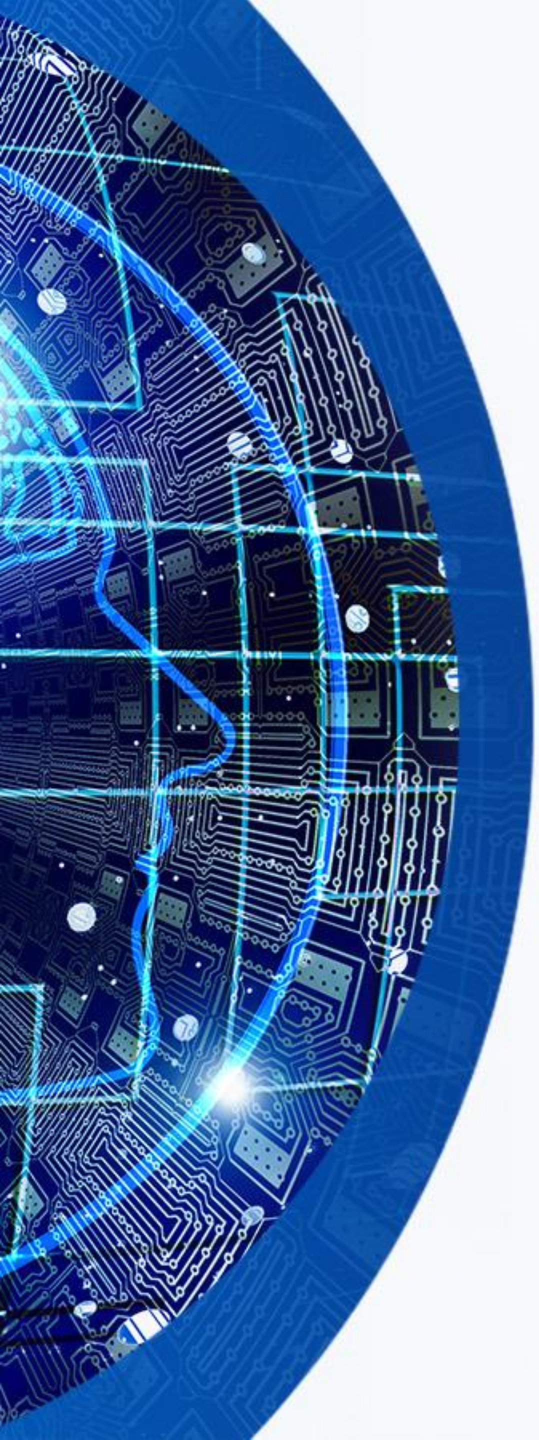


Sentiment Analysis (SA) and Subjectivity

- However, finding opinion sources and monitoring them on the Web can still be a formidable task because there are a large number of diverse sources, and each source may also have a huge volume of **opinionated text** (text with opinions or sentiments).
- In many cases, opinions are hidden in long forum posts and blogs.
- It is difficult for a human reader to find relevant sources, extract related sentences with opinions, read them, summarize them, and organize them into usable forms.

Sentiment Analysis (SA) and Subjectivity

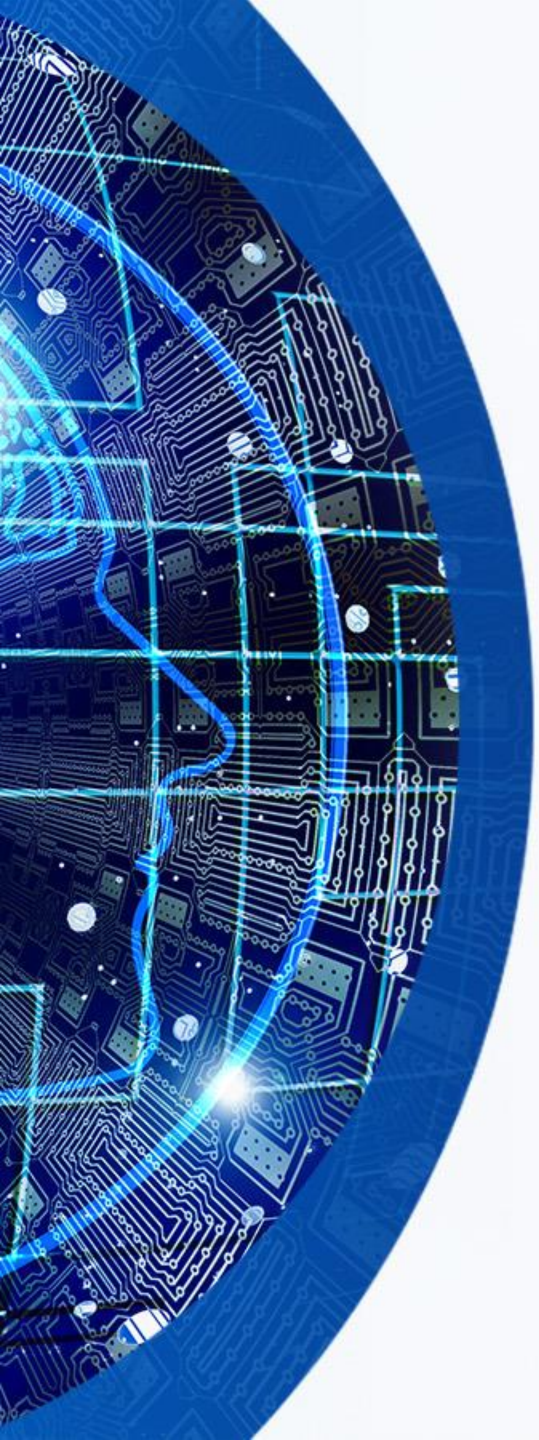
- Thus, automated opinion discovery and summarization systems are needed.
- **Sentiment analysis**, also known as opinion mining, grows out of this need.
- It is a challenging natural language processing or text-mining problem.
- Due to its tremendous value for **practical applications**, there has been an explosive growth of both research in academia and applications in the industry.
- There are now at least 20–30 companies that offer sentiment analysis services in the United States alone.



Sentiment Analysis (SA) and Subjectivity

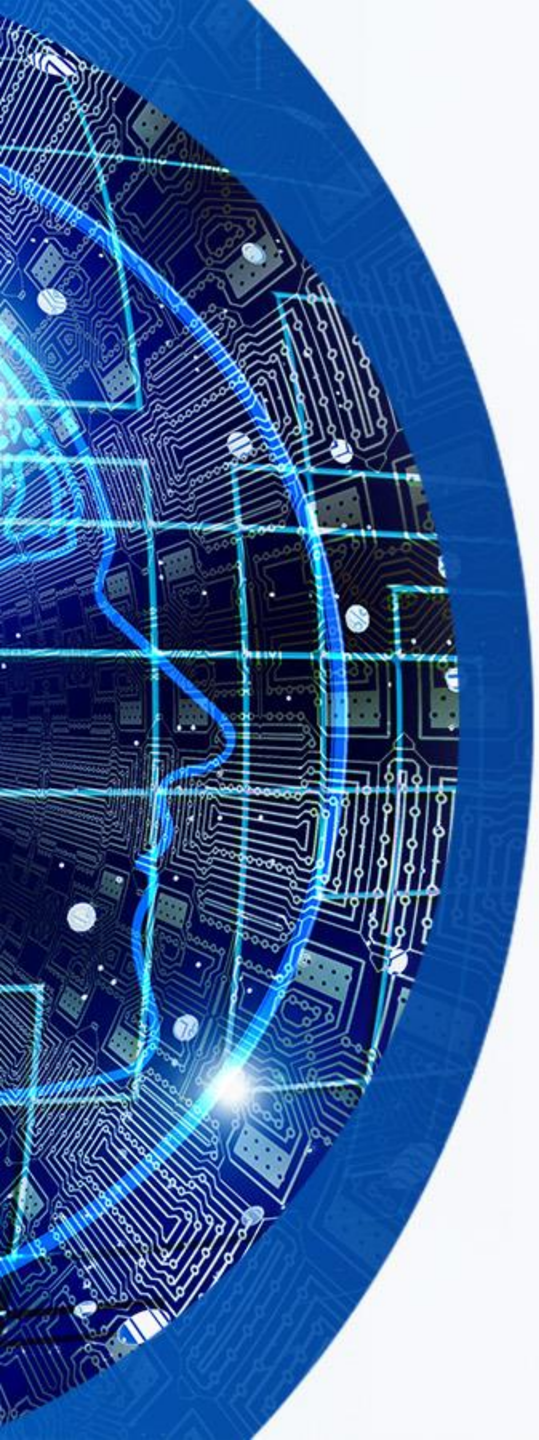
This research field, focuses on the following topics:

- **The problem of sentiment analysis**
- **Sentiment and subjectivity classification**
- **Feature-based sentiment analysis**
- **Sentiment analysis of comparative sentences**
- **Opinion search and retrieval**
- **Opinion spam and utility of opinions**



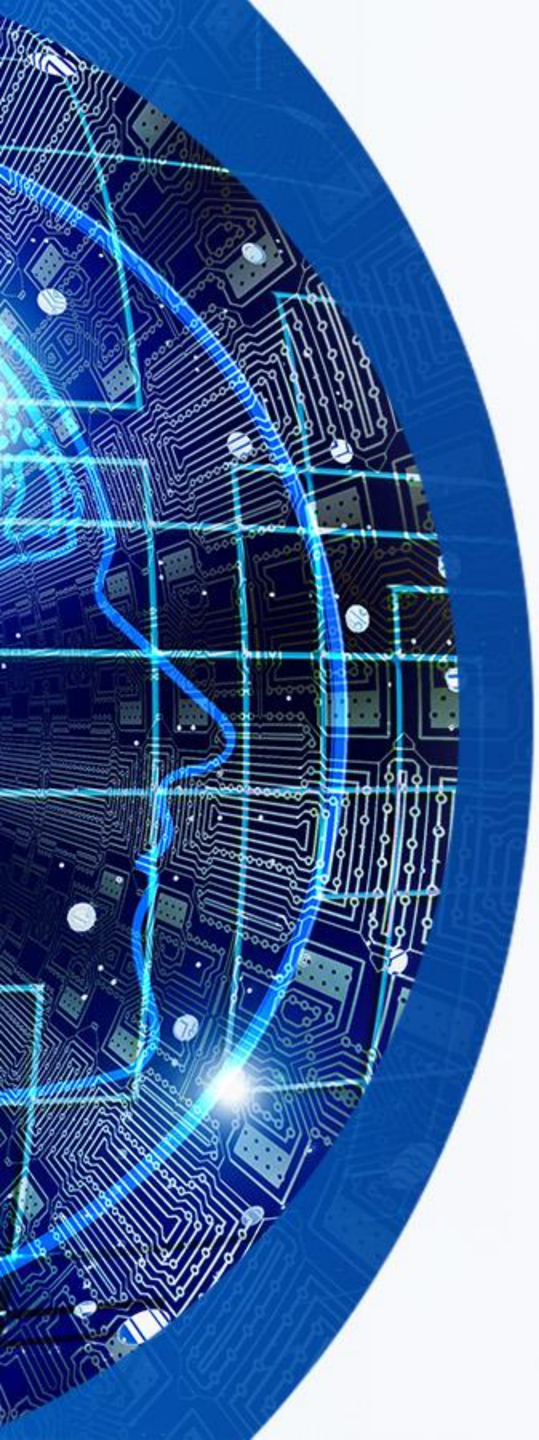
Sentiment Analysis (SA) and Subjectivity

- **The problem of sentiment analysis**: As for any scientific problem, before solving it we need to define or to formalize the problem.
  - ➢ The formulation will introduce the basic definitions, core concepts and issues, subproblems, and target objectives.
  - ➢ It also serves as a common framework to unify different research directions.
  - ➢ From an application point of view, it tells practitioners what the **main tasks** are, their **inputs** and **outputs**, and how the resulting **outputs may be used in practice**



Sentiment Analysis (SA) and Subjectivity

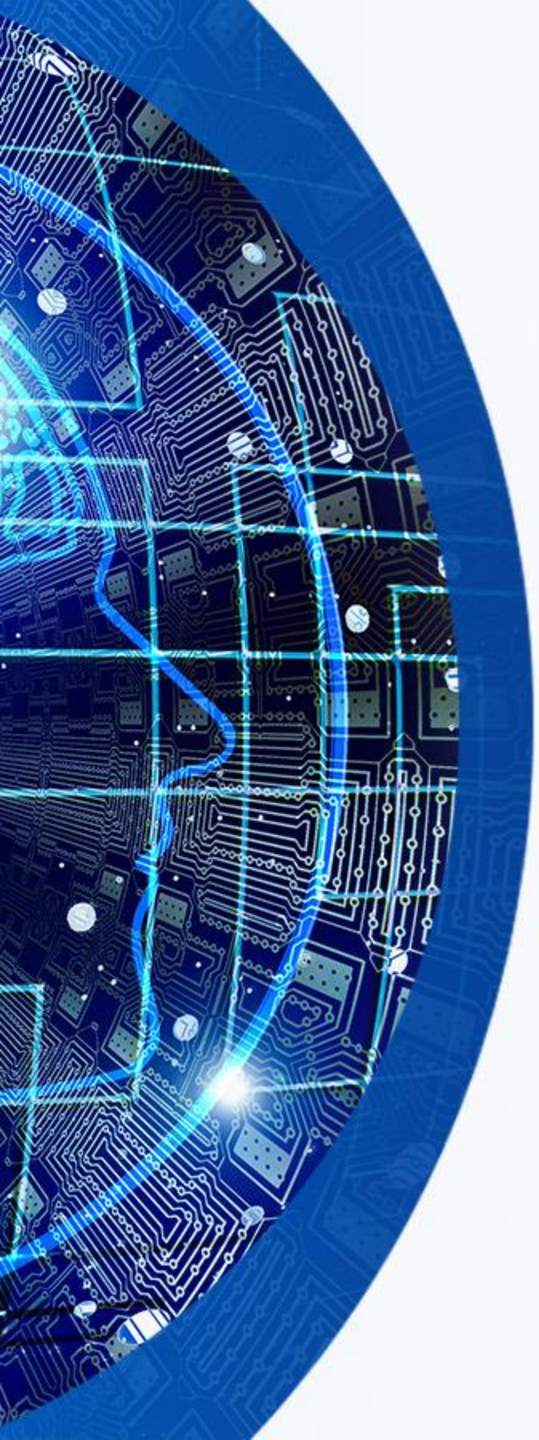
- **Sentiment and subjectivity classification**: It treats sentiment analysis as a text classification problem.
- Two subtopics that have been extensively studied are:
  - (1) classifying an opinionated **document** as expressing a **positive or negative** opinion,
  - (2) classifying a sentence or a clause of the sentence as **subjective** or **objective**, and for a **subjective sentence** or clause classifying it as **expressing a positive, negative,** or neutral opinion.



Sentiment Analysis (SA) and Subjectivity

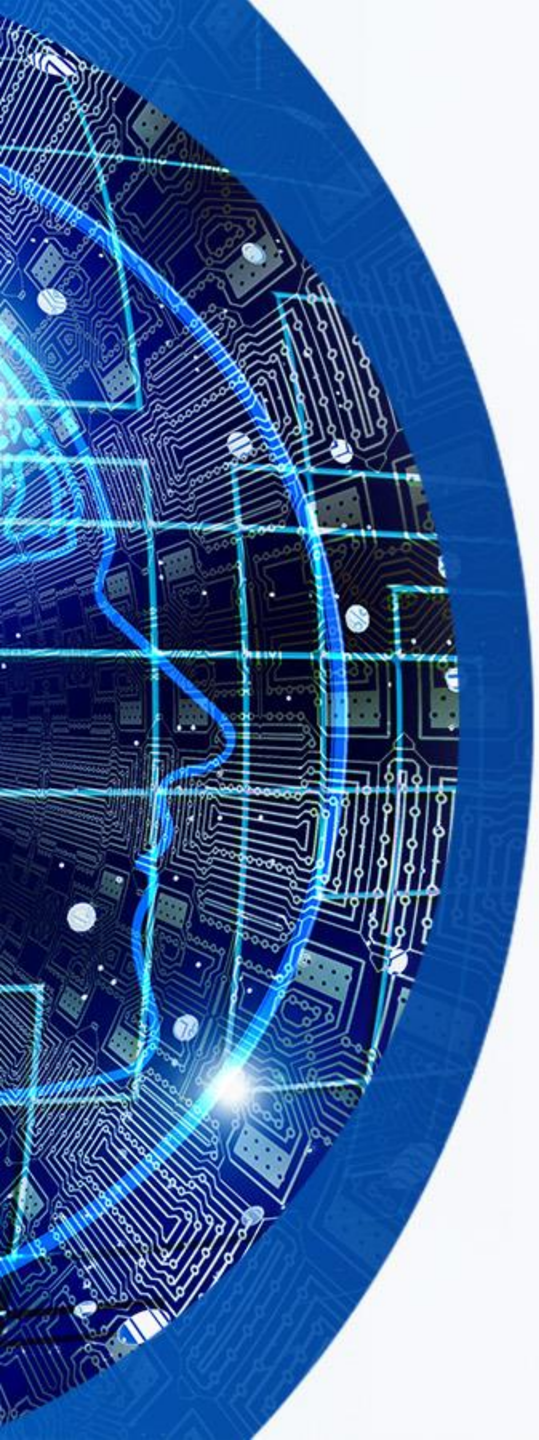
- The first topic, commonly known as sentiment classification or document-level sentiment classification, aims to find the **general sentiment of the author** in an opinionated text.
  For example, given a product review, it determines whether the reviewer is positive or negative about the product.

- The second topic goes to individual sentences to determine whether a sentence expresses an opinion or not (often called **subjectivity classification**), and if so, whether the opinion is positive or negative (called **sentence-level sentiment classification**)



Sentiment Analysis (SA) and Subjectivity

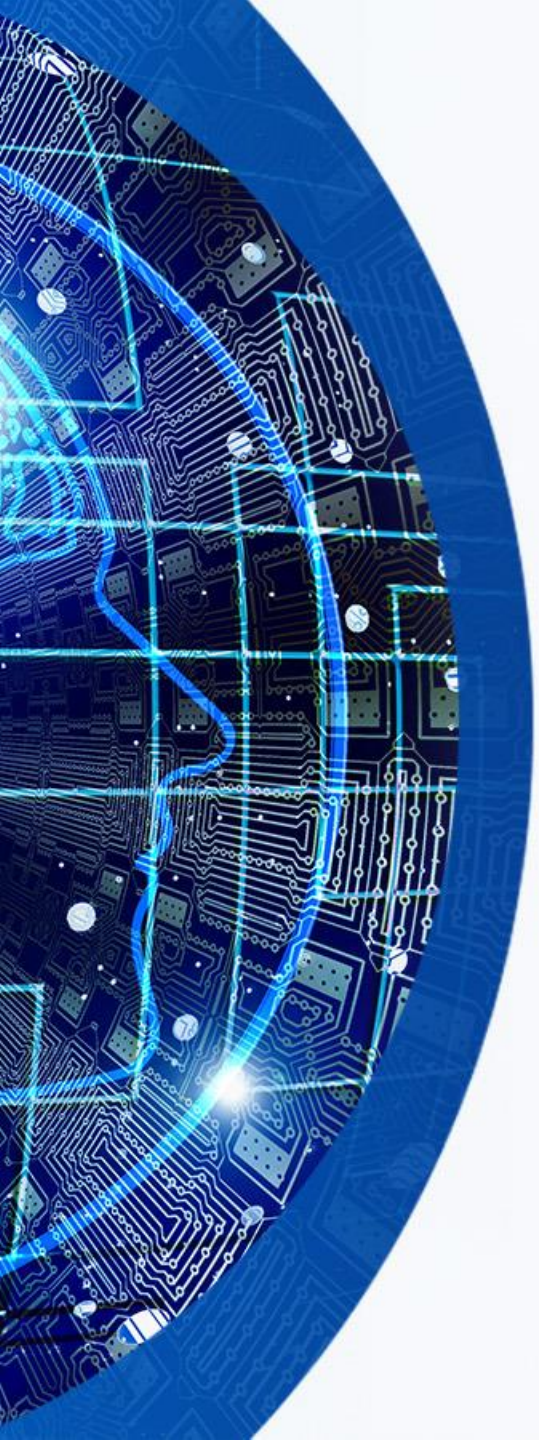
- **Feature-based sentiment analysis**: This model first discovers the targets on which opinions have been expressed in a sentence, and then determines whether the opinions are positive, negative, or neutral.
  - ➢ The targets are objects, and their components, attributes and features.
  - ➢ An object can be a product, service, individual, organization, event, topic, etc.
  - ➢ For instance, in a product review sentence, it identifies product features that have been commented on by the reviewer and determines whether the comments are positive or negative.



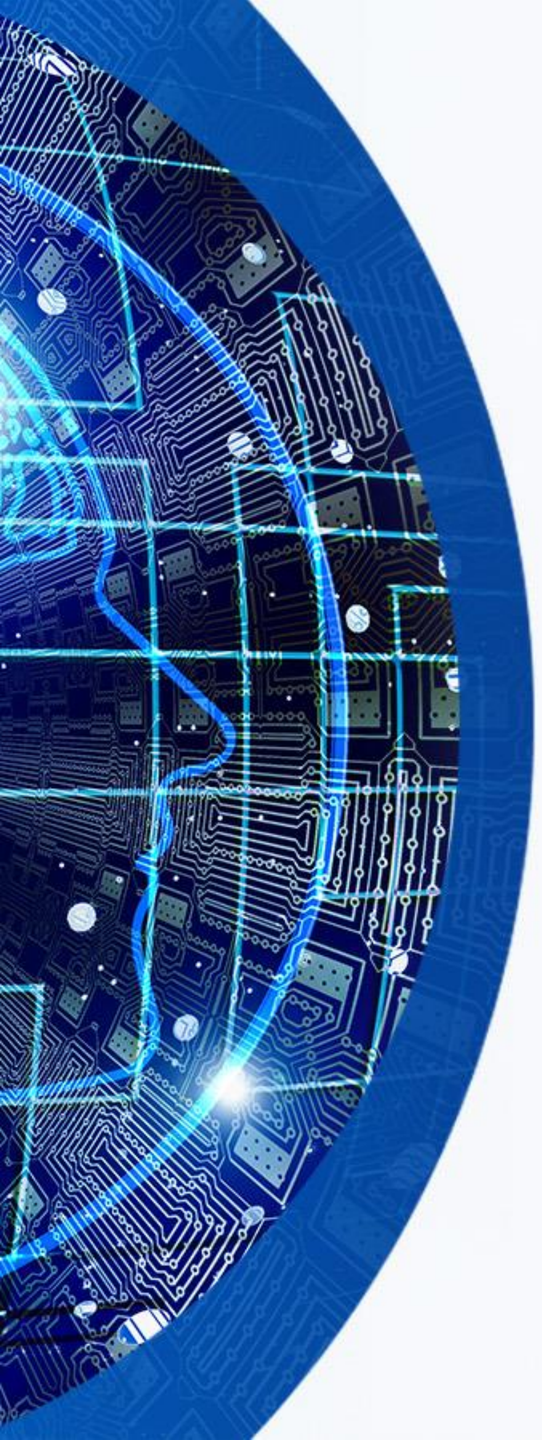
Sentiment Analysis (SA) and Subjectivity

- **Sentiment analysis of comparative sentences**: The evaluation of an object can be done in two main ways, **direct appraisal and comparison**.
  - ➢ Direct appraisal, called direct opinion, gives positive or negative opinions about the object without mentioning any other similar objects.
  - ➢ Comparison means to compare the object with some other similar objects (e.g., competing products).

Sentiment Analysis (SA) and Subjectivity



- For example, “*The picture quality of this camera is poor,*” expresses a direct opinion, while “*The picture quality of Camera-x is better than that of Camera-y.*” expresses a comparison.
- Clearly, it is useful to identify such sentences, extract comparative opinions expressed in them, and determine which objects are preferred by the sentence authors.

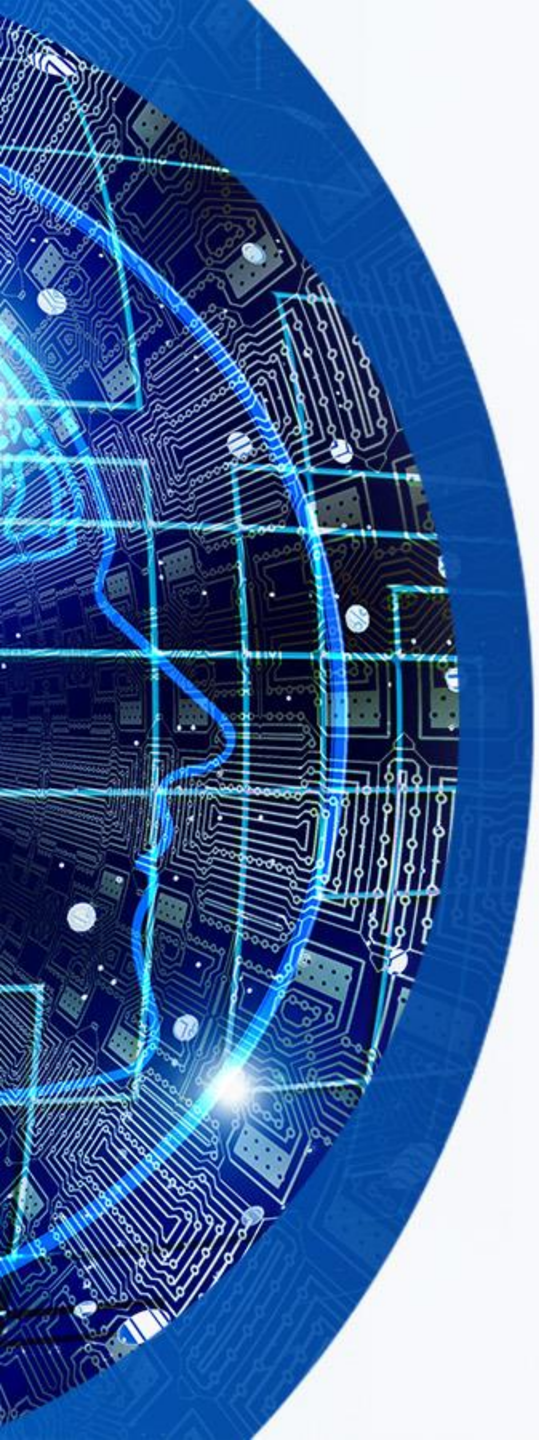


Sentiment Analysis (SA) and Subjectivity

- **Opinion search and retrieval**: Since the general Web search has been so successful in many aspects, it is not hard to imagine that **opinion search** will be very useful as well.
  - ➢ one wants to find positive and negative opinions on the issue from an opinion search engine.
  - ➢ For such a query, two tasks need to be performed:

    (1) retrieving documents or sentences that are relevant to the query, and

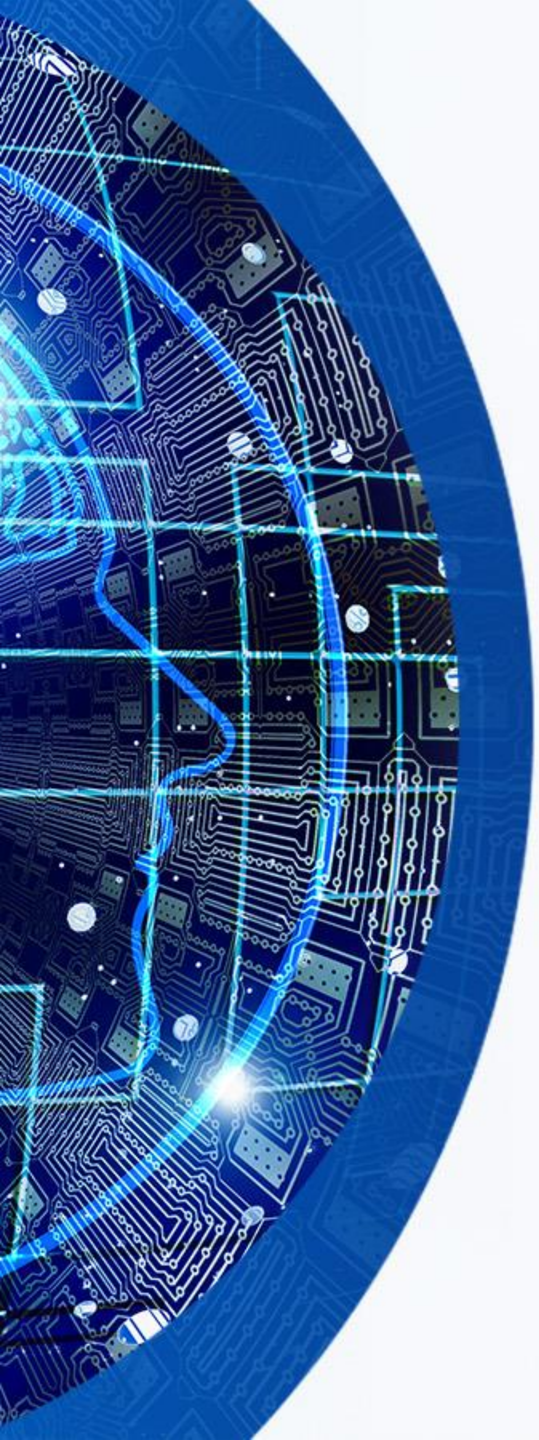
    (2) identifying and ranking opinionated documents or sentences from those retrieved.
  - ➢ Opinion search is thus a combination of IR and sentiment analysis.



Sentiment Analysis (SA) and Subjectivity

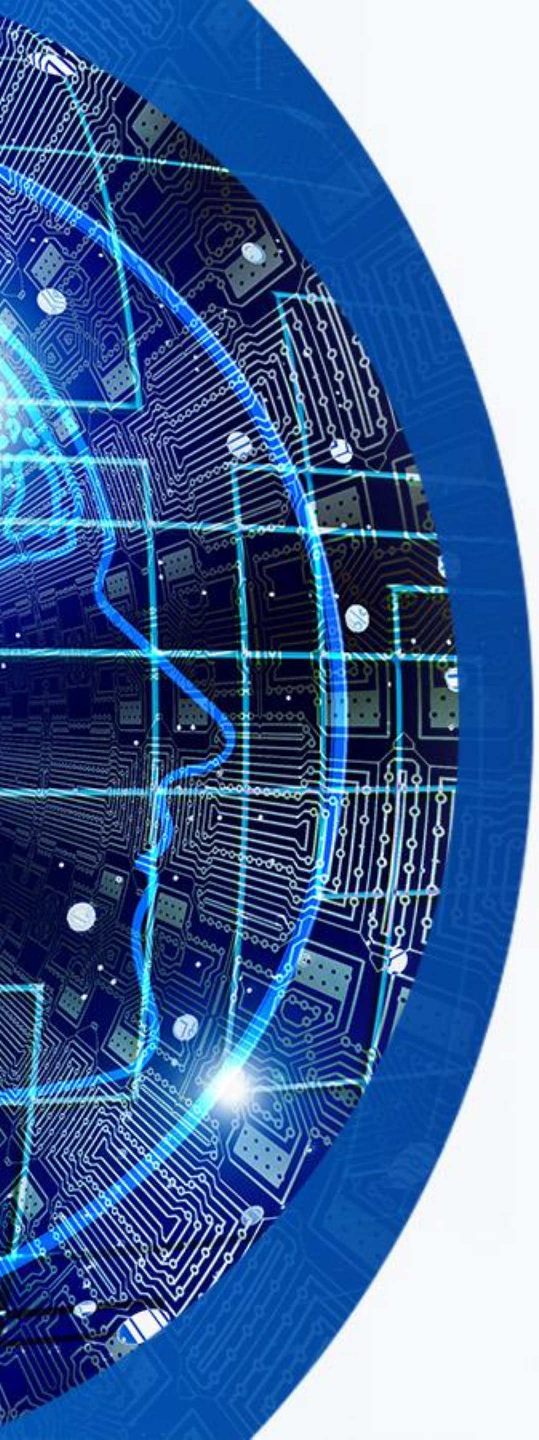
- **Opinion spam and utility of opinions**: As opinions on the Web are important for many applications, it is no surprise that people have started to game the system

  - ➢ **Opinion spam** refers to fake or bogus opinions that try to deliberately mislead readers or automated systems by giving undeserving positive opinions to some target objects in order to promote the objects and/or by giving malicious negative opinions to some other objects in order to damage their reputations.



Sentiment Analysis (SA) and Subjectivity

- Detecting such spam is very important for applications. The utility of opinions refers to the usefulness or quality of opinions.
- Automatically assigning utility values to opinions is useful as opinions can then be ranked based on their utility values.
- With the ranking, the reader can focus on those quality opinions. We should note, however, that spam and utility are different concepts.



Sentiment Analysis (SA) and Subjectivity

- Finally, we conclude the chapter by saying that all the sentiment analysis tasks are very challenging.
- Our understanding and knowledge of the problem and its solution are **still very limited.**
- The main reason is that it is a natural language processing task, and natural language processing **has no easy problems**.
- Another reason may be due to our popular ways of doing research. We probably **relied too much on machine learning algorithms**.
- Some of the most effective machine learning algorithms, produce no human understandable results such that although they may achieve improved accuracy,
- These practical needs and the technical challenges will keep the **field vibrant and lively for years to come.**