Text mining and it is development stages

Maftuna Faxriddin qizi Samiyeva
samiyevamaftuna@gmail.com
Mukhlisa Abdulla qizi Madyarova
madyarovamukhlisa@gmail.com
Tashkent State University of Economics

Abstract: This paper is based on the process of designing text mining. And it illustrates how social mining is increasing in the ICT sector.

Keyword: text mining, text mining methods, NLP, text extraction, spam filtering, text classification, risk management, customer service

INTRODUCTION

Text mining (also known as text analysis), is the process of transforming unstructured text into structured data for easy analysis. Text mining uses natural language processing (NLP), allowing machines to understand the human language and process it automatically.

For businesses, the large amount of data generated every day represents both an opportunity and a challenge. On the one side, data helps companies get smart insights on people's opinions about a product or service. Think about all the potential ideas that you could get from analyzing emails, product reviews, social media posts, customer feedback, support tickets, etc. On the other side, there's the dilemma of how to process all this data. And that's where text mining plays a major role.

Like most things related to Natural Language Processing (NLP), text mining may sound like a hard-to-grasp concept. But the truth is, it doesn't need to be. This guide will go through the basics of text mining, explain its different methods and techniques, and make it simple to understand how it works. You will also learn about the main applications of text mining and how companies can use it to automate many of their processes:

1. Getting started with text mining
2. How does text mining work?
3. Use cases and applications

Text mining is an automatic process that uses natural language processing to extract valuable insights from unstructured text. By transforming data into information that machines can understand, text mining automates the process of classifying texts by sentiment, topic, and intent.

Thanks to text mining, businesses are being able to analyze complex and large sets of data in a simple, fast and effective way. At the same time, companies are taking

advantage of this powerful tool to reduce some of their manual and repetitive tasks, saving their teams precious time and allowing customer support agents to focus on what they do best.

Let's say you need to examine tons of reviews in G2 Crowd to understand what customers are praising or criticizing about your SaaS. A text mining algorithm could help you identify the most popular topics that arise in customer comments, and the way that people feel about them: are the comments positive, negative or neutral? You could also find out the main keywords mentioned by customers regarding a given topic.

In a nutshell, text mining helps companies make the most of their data, which leads to better data-driven business decisions.

At this point you may already be wondering, how does text mining accomplish all of this? The answer takes us directly to the concept of machine learning.

Machine learning is a discipline derived from AI, which focuses on creating algorithms that enable computers to learn tasks based on examples. Machine learning models need to be trained with data, after which they're able to predict with a certain level of accuracy automatically.

When text mining and machine learning are combined, automated text analysis becomes possible.

Going back to our previous example of SaaS reviews, let's say you want to classify those reviews into different topics like UI/UX, Bugs, Pricing or Customer Support. The first thing you'd do is train a topic classifier model, by uploading a set of examples and tagging them manually. After being fed several examples, the model will learn to differentiate topics and start making associations as well as its own predictions. To obtain good levels of accuracy, you should feed your models a large number of examples that are representative of the problem you're trying to solve.

Now that you've learned what text mining is, we'll see how it differentiates from other usual terms, like text analysis and text analytics.

Difference between Text Mining, Text Analysis, and Text Analytics?

Text mining and text analysis are often used as synonyms. Text analytics, however, is a slightly different concept.

So, what's the difference between text mining and text analytics?

In short, they both intend to solve the same problem (automatically analyzing raw text data) by using different techniques. Text mining identifies relevant information within a text and therefore, provides qualitative results. Text analytics, however, focuses on finding patterns and trends across large sets of data, resulting in more quantitative results. Text analytics is usually used to create graphs, tables and other sorts of visual reports.

Text mining combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.

Text analytics, on the other hand, uses results from analyses performed by text mining models, to create graphs and all kinds of data visualizations.

Choosing the right approach depends on what type of information is available. In most cases, both approaches are combined for each analysis, leading to more compelling results.

Methods and Techniques

There are different methods and techniques for text mining. In this section, we'll cover some of the most frequent.

- Basic Methods
- Word frequency

Word frequency can be used to identify the most recurrent terms or concepts in a set of data. Finding out the most mentioned words in unstructured text can be particularly useful when analyzing customer reviews, social media conversations or customer feedback.

For instance, if the words expensive, overpriced and overrated frequently appear on your customer reviews, it may indicate you need to adjust your prices (or your target market!)

Collocation. Collocation refers to a sequence of words that commonly appear near each other. The most common types of collocations are bigrams (a pair of words that are likely to go together, like get started, save time or decision making) and trigrams (a combination of three words, like within walking distance or keep in touch).

Identifying collocations - and counting them as one single word - improves the granularity of the text, allows a better understanding of its semantic structure and, in the end, leads to more accurate text mining results.

Table 1

Here are a few sentences extracted from a set of reviews including the word 'work'

Preceding context	Target	Following context
It saves time and helps teams	**work**	more efficiently.
Some advanced features only	**work**	in one language (English)
It enables us to	**work**	towards better conversion and retention.
We recommend this to several of the small businesses we	**work**	with, and they are all happy with the results.

Concordance. Concordance is used to recognize the particular context or instance in which a word or set of words appears. We all know that the human language can be

ambiguous: the same word can be used in many different contexts. Analyzing the concordance of a word can help understand its exact meaning based on context.

Advanced Methods. Text Classification. Text classification is the process of assigning categories (tags) to unstructured text data. This essential task of Natural Language Processing (NLP) makes it easy to organize and structure complex text, turning it into meaningful data.

Thanks to text classification, businesses can analyze all sorts of information, from emails to support tickets, and obtain valuable insights in a fast and cost-effective way.

Below, we'll refer to some of the most popular tasks of text classification – topic analysis, sentiment analysis, language detection, and intent detection.

Topic Analysis: helps you understand the main themes or subjects of a text, and is one of the main ways of organizing text data. For example, a support ticket saying my online order hasn't arrived, can be classified as Shipping Issues.

Sentiment Analysis: consists of analyzing the emotions that underlie any given text. Suppose you are analyzing a series of reviews about your mobile app. You may find out that the most frequently mentioned topics in those reviews are UI-UX or Ease of Use, but that's not enough information to arrive to any conclusions. Sentiment analysis helps you understand the opinion and feelings in a text, and classify them as positive, negative or neutral. Sentiment analysis has a lot of useful applications in business, from analyzing social media posts to going through reviews or support tickets. In terms of customer support, for instance, you might be able to quickly identify angry customers and prioritize their problems first.

Language Detection: allows you to classify a text based on its language. One of its most useful applications is automatically routing support tickets to the right geographically located team. Automating this task is quite simple and helps teams save valuable time.

Intent Detection: you could use a text classifier to recognize the intentions or the purpose behind a text automatically. This can be particularly useful when analyzing customer conversations. For example, you could sift through different outbound sales email responses and identify the prospects which are interested in your product from the ones that are not, or the ones who want to unsubscribe.

Text Extraction. Text extraction is a text analysis technique that extracts specific pieces of data from a text, like keywords, entity names, addresses, emails, etc. By using text extraction, companies can avoid all the hassle of sorting through their data manually to pull out key information.

Most times, it can be useful to combine text extraction with text classification in the same analysis.

Below, we'll refer to some of the main tasks of text extraction – keyword extraction, named entity recognition and feature extraction.

Keyword Extraction: keywords are the most relevant terms within a text and can be used to summarize its content. Utilizing a keyword extractor allows you to index data to be searched, summarize the content of a text or create tag clouds, among other things.

Named Entity Recognition: allows you to identify and extract the names of companies, organizations or persons from a text.

Feature Extraction: helps identify specific characteristics of a product or service in a set of data. For example, if you are analyzing product descriptions, you could easily extract features like color, brand, model, etc.

Why is Text Mining Important?

Individuals and organizations generate tons of data every day. Stats claim that almost 80% of the existing text data is unstructured, meaning it's not organized in a predefined way, it's not searchable, and it's almost impossible to manage. In other words, it's just not useful.

Being able to organize, categorize and capture relevant information from raw data is a major concern and challenge for companies. Text mining is crucial to this mission.

In a business context, unstructured text data can include emails, social media posts, chats, support tickets, surveys, etc. Sorting through all these types of information manually often results in failure. Not only because it's time-consuming and expensive, but also because it's inaccurate and impossible to scale.

Text mining, however, has proved to be a reliable and cost-effective way to achieve accuracy, scalability and quick response times. Here are some of its main advantages in more detail:

Scalability: with text mining it's possible to analyze large volumes of data in just seconds. By automating specific tasks, companies can save a lot of time that can be used to focus on other tasks. This results in more productive businesses.

Real-time analysis: thanks to text mining, companies can prioritize urgent matters accordingly including, detecting a potential crisis, and discovering product flaws or negative reviews in real time. Why is this so important? Because it allows companies to take quick action.

Consistent Criteria: when working on repetitive, manual tasks people are more likely to make mistakes. They also find it hard to maintain consistency and analyze data subjectively. Let's take tagging, for example. For most teams, adding categories to emails or support tickets is a time-consuming task that often leads to errors and inconsistencies. Automating this task not only saves precious time but also allows more accurate results and assures that a uniform criteria is applied to every ticket.

How Does Text Mining Work?

Text mining helps to analyze large amounts of raw data and find relevant insights. Combined with machine learning, it can create text analysis models that learn to classify or extract specific information based on previous training.

Even though text mining may seem like a complicated matter, it can actually be quite simple to get started with.

The first step to get up and running with text mining is gathering your data. Let's say you want to analyze conversations with users through your company's Intercom live chat. The first you'll need to do is generate a document containing this data.

Data can be internal (interactions through chats, emails, surveys, spreadsheets, databases, etc) or external (information from social media, review sites, news outlets, and any other websites).

The second step is preparing your data. Text mining systems use several NLP techniques - like tokenization, parsing, lemmatization, stemming and stop removal - to build the inputs of your machine learning model.

Then, it's time for the text analysis itself. In this section, we'll explain how the two most common methods for text mining actually work: text classification and text extraction.

Text Classification

Text classification is the process of assigning tags or categories to texts, based on their content.

Thanks to automated text classification it is possible to tag a large set of text data and obtain good results in a very short time, without needing to go through all the hassle of doing it manually. This has exciting applications in different areas.

Rule-based Systems

These type of text classification systems are based on linguistic rules. By rules, we mean human-crafted associations between a specific linguistic pattern and a tag. Once the algorithm is coded with those rules, it can automatically detect the different linguistic structures and assign the corresponding tags.

Rules generally consist of references to syntactic, morphological and lexical patterns. They can also be related to semantic or phonological aspects.

For example, this could be a rule for classifying product descriptions based on the color of a product:

$$(\text{Black} \mid \text{Gray} \mid \text{White} \mid \text{Blue}) \rightarrow \text{Color}$$

In this case, the system will assign the tag COLOR whenever it detects any of the above-mentioned words.

Rule-based systems are easy to understand, as they are developed and improved by humans. However, adding new rules to an algorithm often requires a lot of tests to see if they will affect the predictions of other rules, making the system hard to scale.

Besides, creating complex systems requires specific knowledge on linguistics and of the data you want to analyze.

Text Mining Examples

Text mining is applied throughout a wide variety of industries. Some text mining application examples include:

customer service: businesses can improve the customer experience by gathering customer feedback with a variety of text analytics tools and feedback systems, and text mining and sentiment analysis can help isolate and prioritize customer issues, facilitating real-time responses.

risk management: Text analytics for finance and business helps monitor shifts in sentiments, extracting information from analyst reports in order to derive insights from industry trends.

maintenance: Decision making can be automated by detecting patterns related to problems with product/machinery functionality and the reactive and preventative maintenance processes.

healthcare: Automated information clustering and extraction is crucial for medical research.

spam filtering: Filtering methods may be applied to email text in order to block spam emails and minimize the threat of cyber-attacks on users.

References

1. “Data Mining Concepts and Techniques” Third Edition by Jiawei Han, Micheline Kamber Jian Pei.

2. Fundamentals of Business Intelligence (Data-Centric Systems and Applications) 2015th Edition by Wilfried Grossmann (Author), Stefanie Rinderle-Ma (Author)..

3. “Business Intelligence – Grundlagen und praktische Anwendungen: Eine Einführung in die IT” by Hans-Georg Kemper and Henning Baars.

4. David Loshin Morgan, Kaufman, “Business Intelligence: The Savvy Manager‟s Guide”, Second Edition, 2012 4. Sharda, R., Delen, D. y Turban, E. (2014). Business Intelligence, A Managerial Perspective on Analytics. Boston: Pearson.

5. Bekmuradov A.Sh., Musaliyev A.A., Xashimxodjayev Sh.I. “Axborot biznesi”: O‘quv qo‘llanma. – T.: «IQTISODIYOT», 2019. – 160 b