thematic

UPDATED FEBUARY 2024

5 Text Analytics Approaches: A Comprehensive Review

A review by Dr. Alyona Medelyan, PhD in Natural Language Processing

Contents

Some Text Analytics background....

Throughout my career, I've spoken with many who are living through the pain of analyzing text and trying to find a solution.

Some try to reinvent the wheel by writing their own algorithms from scratch, others believe that ChatGPT can solve everything, others again are stuck with technologies from the late 90's that vendors pitch as "advanced Text Analytics".

I've spent the last 20 years in Natural Language Processing, specifically in the area of making sense of text using algorithms: researching, creating, applying and selling the technology behind it.

My academic research resulted in algorithms used by hundreds of organizations (I'm the author of KEA and Maui). The highlight of my text analytics career was at Google, where I wrote an algorithm that can analyze text in languages I don't speak.

In my role as the CEO of Thematic I've learned a lot about what's available in the market. And most recently, I've been seen the opportunities that Large Language Models brought to this field.

So, it's fair to say, I'm qualified to speak on this topic.

I'll try to be objective in my review, but of **course, I'm biased** because of my position. Happy to discuss this with anyone who is interested in providing feedback.

Alyona Medelyan PhD
CEO and Co-Founder Thematic

Are you receiving more feedback than you could ever read, let alone summarize? Maybe you've used a solution like Medallia, TextIQ or DiscoverXM, or worked with a Python library, or experimented with ChatGPT?

You might be curious to learn more about the methods to analyze free-form textual feedback?

These methods range from simple techniques like word matching to neural networks trained on billions of data points. Here is my summary to break down these methods into 5 key approaches that are commonly used today.

What is text analytics?

Text analytics is the process of extracting meaning out of text. For example, this can be analyzing text written by customers in a customer survey, with the focus on finding common themes and trends. The idea is to be able to examine the customer feedback to inform the business on taking strategic action, in order to improve customer experience.

Organisations can use text analytics software, leveraging machine learning and natural language processing algorithms to find meaning in enormous amounts of text.

How is text analytics used by companies?

To take Thematic as an example, we analyze the free-text feedback submitted by customers and employees through various channels. This was previously difficult to analyze, as companies spend time and resource struggling to do this manually.

Today text analytics helps companies find hidden customer insights and be able to easily answer questions about their existing customer data. In addition, with the help of text analytics software such as Thematic, companies can find recurrent and emerging themes, tracking trends and issues, and create visual reports for managers to track whether they are closing the loop with the end customer.

5 Text Analytics Methods and Examples

Here is my summary to break down these methods into 5 key approaches that are commonly used today.

TEXT ANALYTICS APPROACH 1

Word Spotting

Let's start with **word spotting.** First off, it's not a true analytics thing!

The academic Natural Language Processing community does not register such an approach, and rightly so. In fact, in the academic world, word spotting refers to handwriting recognition (spotting which word a person, a doctor perhaps, has written).

There is also keyword spotting, which focuses on speech processing.

To my knowledge, word spotting is not used in any type of text analytics.

But, I've heard frequently enough about it in meetings to include it in this review. It's loved by DIY analysts and Excel wizards and is a popular approach among many customer insights professionals.

The main idea behind text word spotting is this: If a word appears in text, we can assume that this piece of text is "about" that particular word. For example, if words like "price" or "cost" are mentioned in a review, this means that this review is about "Price".

The beauty of the word spotting approach is its simplicity.

You can implement word spotting in an Excel spreadsheet in less than 10 minutes.
Or, you could write a script in Python or R.

See in the appendix an example of how to do this in Excel.

When word spotting is OK

If you have a dataset with a couple of hundred responses that you only need to analyze once or twice, you can use this approach. If the dataset is small, you can review the results and ensure high accuracy very quickly.

When word spotting fails

As for the downside? Please don't use word spotting:

- If you have any substantial amount of data, more than several hundred responses
- If you won't have time to review and correct the accuracy of each piece of text
- If you need to visualize the results (Excel will hear you swearing)
- If you need to share the results with your colleagues
- If you need to maintain the data consistently over time

There are also many other disadvantages to DIY word spotting, that we'll discuss in the next post. I'll also talk about what actually does work and is a good approach.

TEXT ANALYTICS APPROACH 2

Manual Rules

The Manual Rules approach is closely related to word spotting. Both approaches operate on the same principle of creating a match pattern,, and is a recognized text analytics approach. Both approaches follow a principle of creating a match pattern but Manual Rules covers more complex patterns.

For example, a manual rule could involve the use of regular expressions – something you can't easily implement in Excel. Here is a rule for assigning the category **"Staff Knowledge"** from a popular enterprise solution Medallia:

"Staff knowledge"

i. knowledge

ii. knowledge NEAR where WITH 1 words

iii. knowledge CONNECTED TO products

Manual rules are used by the majority of text analytics software and customer experience management providers who sell text analytics as part of the platform. Their software interface makes it easy to create and manage such rules. They may also sell professional services to help with the creation and maintenance of these rules.

The best thing about Manual Rules is that they can be understood by a person. They are explainable, and therefore can be tweaked and adjusted when needed.

But the bottom line is that creating these rules takes a lot of effort. You also need to ensure that they are accurate and maintain them over time.

To get you started, some companies come with pre-packaged rules, already organized into a taxonomy. For example, they would have a category "Price", with hundreds of words and phrases already pre-set, and underneath they might have sub-categories such as "Cheap" and "Expensive".

They may also have specific categories setup for certain industries, e.g. banks. And if you are a bank, you just need to add your product names into this taxonomy, and you're good to go.

The benefit of this approach is that once set up, you can run millions of feedback pieces and get a good overview of the core categories mentioned in the text.

But, there are plenty of disadvantages for this approach, and in fact any manual rules and word spotting technique:

1. Multiple word meanings make it hard to create rules

The most common reason why rules fail stems from **polysemy**, when the same word can have different meanings:

It's hard to find a rule that works well.

Friendly OR friendliness	Staff friendliness	
I was impressed by how **friendly** the person on the other end of the line was	Staff friendliness	✓
The lady who helped me was **friendly**	Staff friendliness	✓
Friendliness of staff	Staff friendliness	✓
Your website is very user **friendly**	Staff friendliness	✗
The young man on the phone was very **friendly**	Other	✗

2. Mentioned word ≠ core topic

Just because a word or a phrase is mentioned in text, it doesn't always mean that the text is about that topic.

For example, when a customer is explaining the situation that leads to an issue: "My credit card got declined and the cashier was super helpful, waiting patiently while I searched for cash in my bag." This comment is not about credit cards or cash, it's about the behavior of the staff.

3. Rules cannot capture sentiment

Knowing the general category alone isn't enough.
How do people think about "Price", are they happy or not?

	Guessed sentiment based on "great"	Actual sentiment
My coffee was great	Positive	Positive
My coffee was awful	Negative	Negative
My coffee was not great	Positive	Negative
My coffee was not that great	Positive	Neutral?
I did not think my coffee was great	Positive	Negative
I did not expect my coffee to be this great	Positive	Positive
I was dissapointed with the quailty of the coffee	Negative	Negative
I wasn't dissapointed with the quailty of the coffee	Negative	Positive

Capturing sentiment with manually pre-set rules is impossible. People often do not realize how diverse and varied our language is.

So, a sub-category like "expensive" is actually extremely difficult to model. A person could say something like "I did not think this product was expensive". To categorize this comment into a category like "good price", you would need a complex algorithm to detect negation and its scope. A simple regular expression won't cut it.

4. Taxonomies don't exist for software products and many other businesses

The pre-set taxonomies with rules won't exist for non-standard products or services. This is particularly problematic for the software industry, where each product is unique and the customer feedback talks about very specific issues.

5. Not everyone can maintain rules

In any industry, even if you have a working rule-based taxonomy, someone with good linguistic knowledge would need to constantly maintain the rules to make sure all of the feedback is categorized accurately. This person would need to constantly scan for new expressions that people create so easily on the fly, and for any emerging themes that weren't considered previously. It's a never-ending process which is highly expensive.

And yet, despite these disadvantages, this approach is the most widely used commercial application of Text Analytics, with its roots in the 90s, and no clear path for fixing these issues.

So, are Manual Rules good enough?

My answer to this is **No**. Most people who use Manual Rules are dissatisfied with the time required to set up a solution, with the costs to maintain it.

TEXT ANALYTICS APPROACH 3

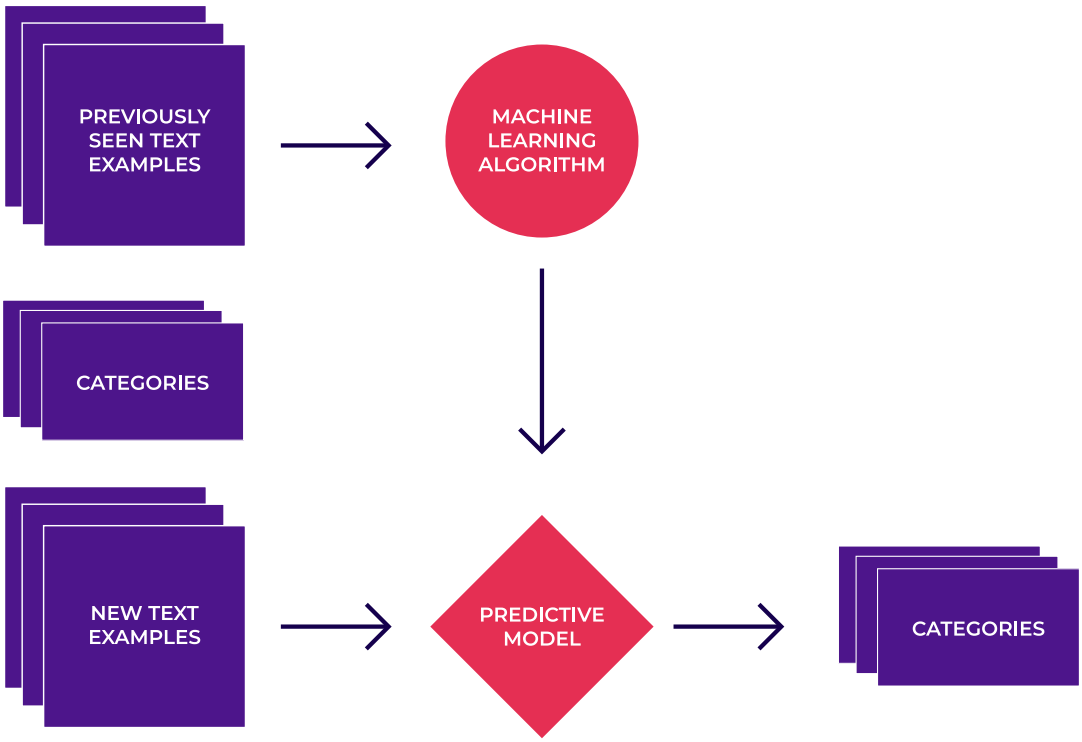
Text Categorization

Let's bring some clarity to the messy subject of **Advanced Text Analytics,** the way it's pitched by various vendors and data scientists.

Here, we'll be looking at **Text Categorization,** the first of the three approaches that are actually automated and use algorithms.

What is text categorization?

This approach is powered by machine learning. The basic idea is that a machine learning algorithm (there are many) analyzes previously manually categorized examples (the training data) and figures out the rules for categorizing new examples. It's a supervised approach.



The beauty of text categorization is that you simply need to provide examples, no manual creation of patterns or rules needed, unlike in the two previous approaches.

Another advantage of text categorization is that, theoretically, it should be able to capture the relative importance of a word occurrence in text. Let's revisit the example from earlier posts. A customer may be explaining the situation that leads to an issue:

> "My credit card got declined and the cashier was super helpful, waiting patiently while I searched for cash in my bag."

This comment is not about credit cards or cash, it's about the behaviour of the staff. The theme "credit card" mentioned in the comment isn't important, but "helpfulness" and "patience" is. A text categorization approach can capture it with the right training.

It all comes down seeing similar examples in the training data.

Near perfect accuracy... but only with the right training data

There are academic research papers that show that text categorization can achieve near perfect accuracy. Deep Learning algorithms are even more powerful than the old naïve ones (one older algorithm is actually called Naïve Bayes).

And yet, all researchers agree that **the algorithm isn't as important as the training data.**

The quality and the amount of the training data is the deciding factor in how successful this approach is for dealing with feedback. So, how much is enough? Well, it depends on the number of categories and the algorithm used to create a categorization model.

The more categories you have and the more closely related they are, the more training data is needed to help the algorithm to differentiate between them.

Some solutions that rely on text categorization provide tools that make it easy for people to train the algorithms, so that they get better over time.

But do you have time to wait for the algorithm to get better, or do you need to act on customer feedback today?

Four issues with text categorization

Apart from needing to train the algorithm, here are four other problems with using text categorization for analyzing people's feedback:

1. You won't notice emerging themes

You will only learn insights about categories that you trained for and will miss the unknown unknowns. This is the same disadvantage as manual rules and word spotting has: The need to continuously monitor the incoming feedback for emerging themes, and mis-categorized items.

2. Lack of transparency

While the algorithm gets better over time, it is impossible to understand why it works the way it works and therefore easily tweak the results. Qualitative researchers have told me that the lack of transparency is the main reason why text categorizatio did not take off in their world. For example, if there is suddenly poor accuracy on differentiating between two themes "wait time to install fiber" and "wait time on the phone to setup fiber", how much training data does one need to add, until the algorithm stops making these mistakes?

3. Preparing and managing training data is hard

The lack of training data is a real issue. It's hard to start from scratch and most companies don't have enough or accurate enough data to train the algorithms. In fact, companies always overestimate how much training data they have, which makes implementation fall below expectations. And finally, if you need to refine one specific category, you will need to re-label all of the data from scratch.

4. Re-training for each new dataset

Transferability can be really problematic! Imagine you have a working text categorization solution for one of your departments, e.g. support, and now want to analyze feedback that comes through customer surveys, like NPS or CSAT. Again, you would need to re-train the algorithm.

I just got of the phone with a subject matter expert on survey analysis, who told me this story: A team of data scientists spent many months and created a solution that she ultimately had to dismiss due to lack of accuracy. The company did not have time to wait for the algorithm to get better over time.

TEXT ANALYTICS APPROACH 4

Large Language Models

What is a language model?

A language model is created by analyzing a large body of text. The model records word frequencies: on their own, and in different contexts. This means that a language model can determine how similar two words are. It can also predict which word should follow the next one. It can also generate human-like text.

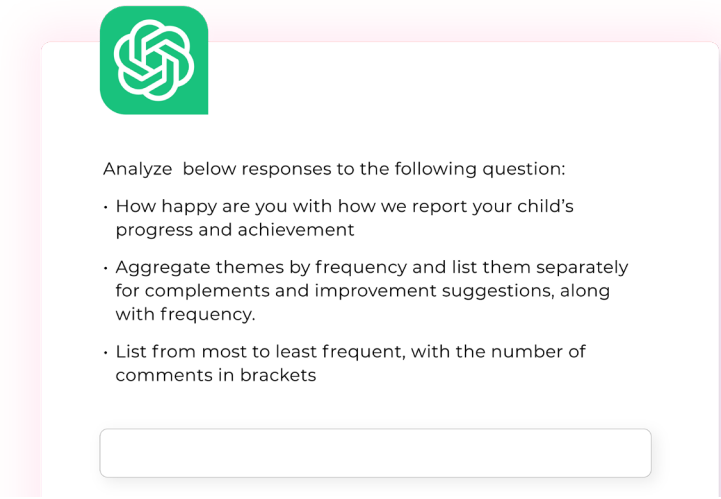
The quality of the language model depends on the amount of data it has seen and iterated over. Large language models (or LLMs) have been trained on billions of data points.

GPT-3 was the first publicly known large language model. Its creators, OpenAI, have now shipped GPT-3.5 and GPT-4, both significant improvements over each other.

Other companies have shipped similar LLMs, e.g. Claude by Anthropic, LLaMa by Meta, PaLM by Google etc.

To analyze text data, such as customer feedback, you need to write a prompt to guide the analysis. The quality of the output greatly depends on this prompt. It needs to provide context and be specific.

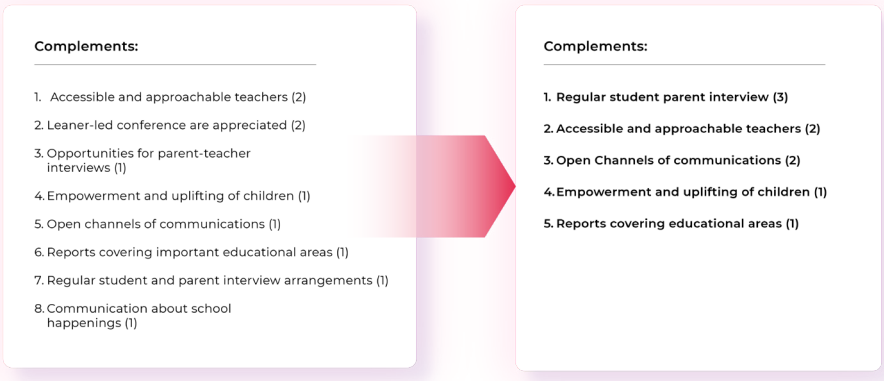
Here's an example of how I instructed GPT-3.5 to analyze a school feedback dataset with 100 parent comments:



You might get satisfactory results straight away!

But more common than not, you will need to solve for errors since, unfortunately, LLMs will create duplicates of the same themes.

Here's an example of me fixing errors resulting from the above prompt.



Read the full article on [how to analyze feedback using ChatGPT here.]

What's great about Large Language Models:

While the jury is out whether LLMs truly understand text, this is the closest we ever gotten in the field of AI. An LLM can correctly interpret acronyms, resolve pronouns, analyze and summarize specialized text at a level that's often above an average human.

Forget manual rules, training categorization models on manual examples! LLMs can make sense of customer feedback at scale more efficiently than any of these methods. They are fast, no cleaning or training needed.

The only challenge is in writing an accurate prompt and tweaking the resulting analysis for reporting.

Where LLMs fall short:

You can very efficiently analyze a small dataset of 100-500 rows of feedback, and interact with the AI via an interface like ChatGPT to resolve any issues. But analyzing large amounts of feedback requires quite a bit of engineering.

Once there is a lot more data, LLMs start to hallucinate. They pull out insights that aren't real. This was especially a big issue with earlier LLMs. But even newer models like to latch on to proper names to sound knowledgeable.

We also found that LLMs struggle to manage more than 20 themes. They end up creating duplicate themes or often miss themes that are present in the data.

As a result, you need some method of verifying the analysis. But once there are huge volumes of data, LLMs become black boxes that make this impossible.

In order to analyze feedback at scale and over a period of time, LLMs alone won't suffice. You'll need to engineer a solution that can manage hallucinations, discovery of emerging themes and tracking these over time.

TEXT ANALYTICS APPROACH 5

Thematic Analysis

(plus our secret sauce on how to make it work even better)

All of the former approaches mentioned have disadvantages. In the best case, you'll get OK results only after spending many months setting things up. And you may miss out on the unknown unknowns.

The cost of acting late or missing out on crucial insights is huge! It can lead to losing customers and stagnant growth. On the flip-side, when companies act on feedback they grow faster. According to the American Customer Satisfaction Index study comparing leaders vs. S&P 500, companies that invest into insights get 4x better stock returns.

When it comes to customer feedback, three things matter:

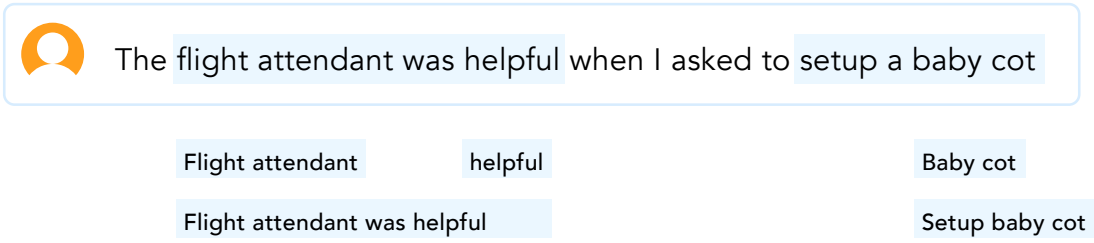
1. Accurate, specific and actionable analysis
2. Ability to see emerging themes fast, without the need of setting things up
3. Transparency in how results are created, to bring in domain expertise and common sense knowledge

In my research, I've learned that the only approach that can achieve all three requirements is Thematic Analysis, combined with an interface for easily editing the results.

Thematic Analysis: How it works

Thematic Analysis approaches extract themes from text, rather than categorize text.

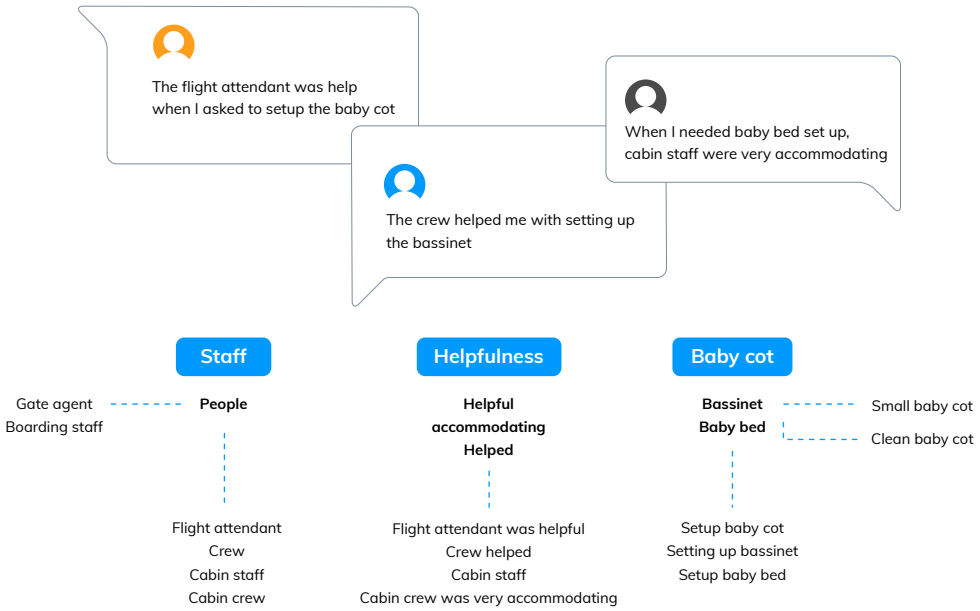
In other words, it's a bottom-up analysis. Given a piece of feedback such as "The flight attendant was helpful when I asked to set up a baby cot", they would extract themes such as "flight attendant", "flight attendant was helpful", "helpful", "asked to set up a baby cot", and "baby cot".



These are all meaningful phrases that can potentially be insightful when analyzing the entire dataset.

However, the most crucial step in a Thematic Analysis approach is merging phrases that are similar into themes and organizing them in a way that's easy for people to review and edit. We achieve this by using our custom word embeddings implementation, but there are different ways to achieve this. We also lean in heavily on LLMs to both merge and organize themes into a taxonomy of base themes and sub-themes.

For example, here is how three people talk about the same thing, and how we at Thematic group the results into themes and sub-themes:



Advantages and disadvantages of Thematic Analysis

The advantage of Thematic Analysis is that this approach is unsupervised, meaning that you don't need to set up these categories in advance, don't need to train the algorithm, and therefore can easily capture the unknown unknowns.

The disadvantages of this approach are that it's difficult to implement correctly if you were to do it in house. A perfect approach must be able to merge and organize themes in a meaningful way, producing a set of themes that are not too generic and not too large. Ideally, the themes must capture at least 80% of verbatims (people's comments). And the themes extraction must handle complex negation clauses, e.g. "I did not think this was a good coffee".

Who does Thematic Analysis?

Some of the established bigger players have implemented Thematic Analysis to enhance their Manual Rules approaches but tend to produce a laundry list of terms that are hard to review.

Traditional Text Analytics APIs designed by NLP experts also use this approach. However, they are rarely designed with customer feedback in mind and try to solve this problem in a generic way. For example, when we tested Google and Microsoft's APIs we found that they aren't grouping themes out of the box.

As a result, only 20 to 40% of feedback is linked to top 10 themes: only when there are strong similarities in how people talk about specific things. The vast majority of feedback is uncategorized meaning that you can't slice the data for deeper insights.

At Thematic, we have developed a Thematic Analysis approach that can easily analyze feedback from customers of pizza delivery services, music app creators, real estate brokers and many more. We achieved this by focusing on a specific type of text: customer feedback, unlike NLP APIs or LLMs that are designed to work on any type of text. We have implemented complex negation algorithms that separate positive from negative themes, to provide better insight.

Our secret sauce: Human in the loop

Each dataset, and sometimes even each survey question, gets its own set of themes, and by using our Themes Editor, insights professionals can refine the themes to suit their business. For example, Thematic might find themes such as "fast delivery", "quick and easy", "an hour wait", "slow service", "delays in delivery" and group them under "speed of service". One insight professional might re-group these into "slow" and "fast" under "speed of service", another into "fast service" > "quick and easy", and "slow service" -> "an hour wait", "delays in delivery". It's a subjective task.

I believe more and more companies will discover Thematic Analysis, because unlike all other approaches, it's a transparent and deep analysis that does not require training data or time for crafting manual rules.

What are your thoughts?

Thematic Text Analytics Cheat Sheet

Approach	Thematic Analysis	Text Categorization	Large Language models	Manual Rules and Taxonomies
How it works	Themes are extracted from text, similar ones merged.	Categories trained on pre-categorized data.	Write a prompt to interpret the data.	Manually crafted and maintained rules.
Best fit	Companies in any industries who have at least a few hundred pieces of feedback per month.	Companies who manually and consistently tag feedback and don't make major changes to their offering.	Companies with expert AI team for a one- off analysis for quick insights.	Companies in industries that don't change their offering.
Data required	A few hundred pieces of feedback a month.	At least a few hundred feedback pieces per category.	Any amount, but best for small data sets that fit into a single prompt.	Any amount.
Advantages	Does not require training data. Captures unknowns. Easy to understand. Can be accurate and capture context.	Can be highly accurate and capture context.	No training required.	Easy to understand.
Disadvantages	Difficult to implement correctly, e.g. negation must be handled accurately.	Requires training data. Can't capture unknowns. Difficult to tweak.	Difficult to tweak the analysis.	Requires a lot of time to create. Can't capture unknowns or sentiment.
Effort to setup	Days	Months	Days to weeks	Months
Effort to maintain	Anyone can maintain. 1-2h per week for Thematic to review.	Not much, if categories don't change.	Depends on the in house skill set.	Professional services: 1 person 1 day a week.
Accuracy	Good to Very High	Good to Very High	Moderate - Very High	Moderate to Good
Transparency	Very Good	Poor	Poor	Good

APPENDIX

How to analyze your data in Excel

How to build a Text Analytics solution in 10 minutes

You can type in a formula, like this one, in Excel to categorize comments into “Billing”, “Pricing” and “Ease of use”:

```
=IF(ISNUMBER(SEARCH("bill",B2)),"BILLING",IF(ISNUMBER(SEARCH("cost",B2)),"PRICE",IF(ISNUMBER(SEARCH("eas",B2)),"EASE OF USE")))
```

And voilà!

Here it is applied to a Net Promoter Score survey where column B contains open- ended answers to questions “Why did you give us this score”:

	A	B	C
1	NPS Category	Comment	Category
2	3	Always cost	PRICE
3	1	Because I find my gas bills are crazy, even in summer.	BILLING
4	10	Very easy to understand web site	EASE OF USE
5	8	easy to pay and different payment options	EASE OF USE
6	10	Everything was easy to do and I had no problems!	EASE OF USE
7	8	it was easy	EASE OF USE
8	9	It's easy and fast	EASE OF USE

It probably took me less than 10 minutes to create this, and the result is so encouraging!
But wait...

Everyone loves simplicity. But in this case, simplicity sucks.

Various issues can easily crop up with this approach. Here, I've annotated them for you.

	A	B	C	D
1	NPS Category	Comment	Category	Notes
2	3	Always cost	PRICE	
3	1	Because I find my gas bills are crazy, even in summer.	BILLING	Should be "Price"
4	10	Very easy to understand web site	EASE OF USE	Missed the Website aspect
5	8	easy to pay and different payment options	EASE OF USE	Missed the Ease of Payment aspect
6	10	Everything was easy to do and I had no problems!	EASE OF USE	
7	8	it was easy	EASE OF USE	
8	9	It's easy and fast	EASE OF USE	Missed "Fast"

Out of 7 comments, here only 3 were categorized correctly.
"Billing" is actually about **"Price"**, and three other comments missed additional themes.

Would you bet your customer insights on something that's at best 50% accurate?

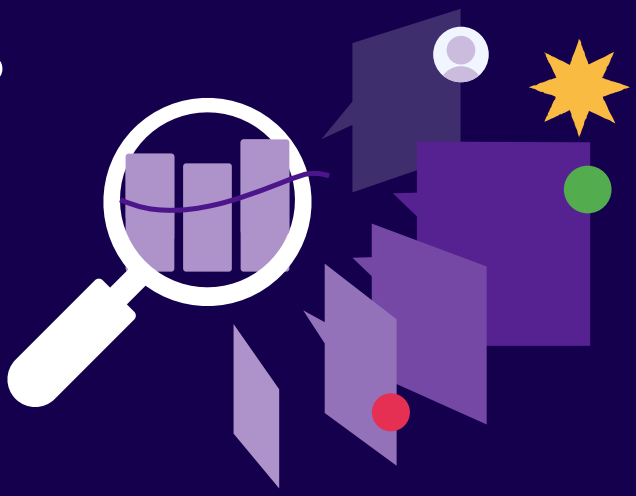
thematic

Want to spot customer issues before it's too late?

Thematic allows you to turn customer feedback into actionable insights. We provide a strategic, in-depth analysis of your customer feedback through AI text analytics.

Book a consult with one of our team - we'd be thrilled to show you how Thematic works!

Talk to one of our experts



getthematic.com