Data Science and its Practical Applications in Multilingual Content

By Hannes Ben

Locaria

Locaria

Contents

Executive summary

Data is everywhere and used by all of us all of the time. Many of us are not even aware of it, but every decision that we make is based on experiences and information that we assimilate, compare, and then decide are best suited to our personal goals. We make unconscious statistical decisions regarding what may have a better impact on what we want to achieve, and then continue throughout our life and work to enrich our database of information to further deepen our understanding and strengthen our confidence in our decisions and actions.

Therefore, in reality, data and the science around data, otherwise called 'data science', are not abstract concepts and should not be difficult to grasp.

Within this paper, we explore the symbiosis of data science and linguistics, including practical applications of data science in language services, particularly multilingual research, production and optimisation in the digital content space.

We support these insights with examples of analysis to understand the performance of data science in localisation, such as incrementality testing and Performance Linguistics™.

Introduction

Over the last 10 years the language, digital marketing, and creative advertising industries have all started to assemble teams focusing specifically on gathering large volumes of data relevant to their work, querying that data, and devising tests to prove concepts. Some have been more successful than others and managed to convert complex ideas and assumptions into meaningful measurement frameworks that have resulted in efficient and effective marketing campaigns.

The language service industry understood the importance of data science to their field a while ago and either assembled their own specialised teams or started to partner with relevant technological and data firms to manage the ever-growing volume of translation required. Particularly with the growth in digital, it has become much easier to reach audiences around the world as long as your owned, earned and paid media content is properly localised. However, with this huge amount of content creation and translation required, often daily, traditional and purely human-based translation processes could not keep up. Originally, all machine translation was based on rule-based machine translation (RBMT), which was first trialled in the 1950s. At that time, computers used bilingual dictionaries and linguistic rules to tackle grammatical challenges, but that never produced even remotely acceptable results. Then, in the 1990s, statistical machine translation quickly gained popularity, as it produced significantly better results by using large amounts of professional translated parallel data (source and target language aligned in one database). Finally, from 2010 onwards, deep neural networks were introduced and allowed for surprisingly accurate translation output. For a more detailed overview of SMT and NMT with a more technical explanation as to their workings, please refer to a great article on them on the Towards Data Science website[1].

It was in the last 10 years that engineers, data scientists, and programmers became truly excited about machine translation and other technologies involving 'natural language processing' (NLP), as Python (with its wide range of relevant coding libraries) has made it faster and easier to handle complex statistical modelling. The excitement surrounding being able to produce acceptable translation through

the use of computer science, and specifically the ongoing development of neural networks and NLP processes, has started to draw some of the best talent towards the translation industry[2].

This need for data-driven approaches is not purely driven by the agency's interest or by the industry's advances in technology; in fact, it is the clients who demand from their strategists, planners and analysts that they base decisions on proven facts and statistically significant results and take action with great confidence for success.

In the language industry or creative advertising, most content creations, while following strict brand style and tone-of-voice guidelines, will still often be largely subjective and ultimately decided by a smaller group of people. There is often little time with which to pressure-test ideas with a large cohort, and even if conducted, the cohort is unlikely to be sufficiently representative of the total volume of people targeted.

It is exactly in those industries where you often encounter heated discussions surrounding how much data should be used to inform your ideas and decisions. Does it harm creativity? Are machines going to take over? Does my content lose the human touch? Does it come across as being too automated?

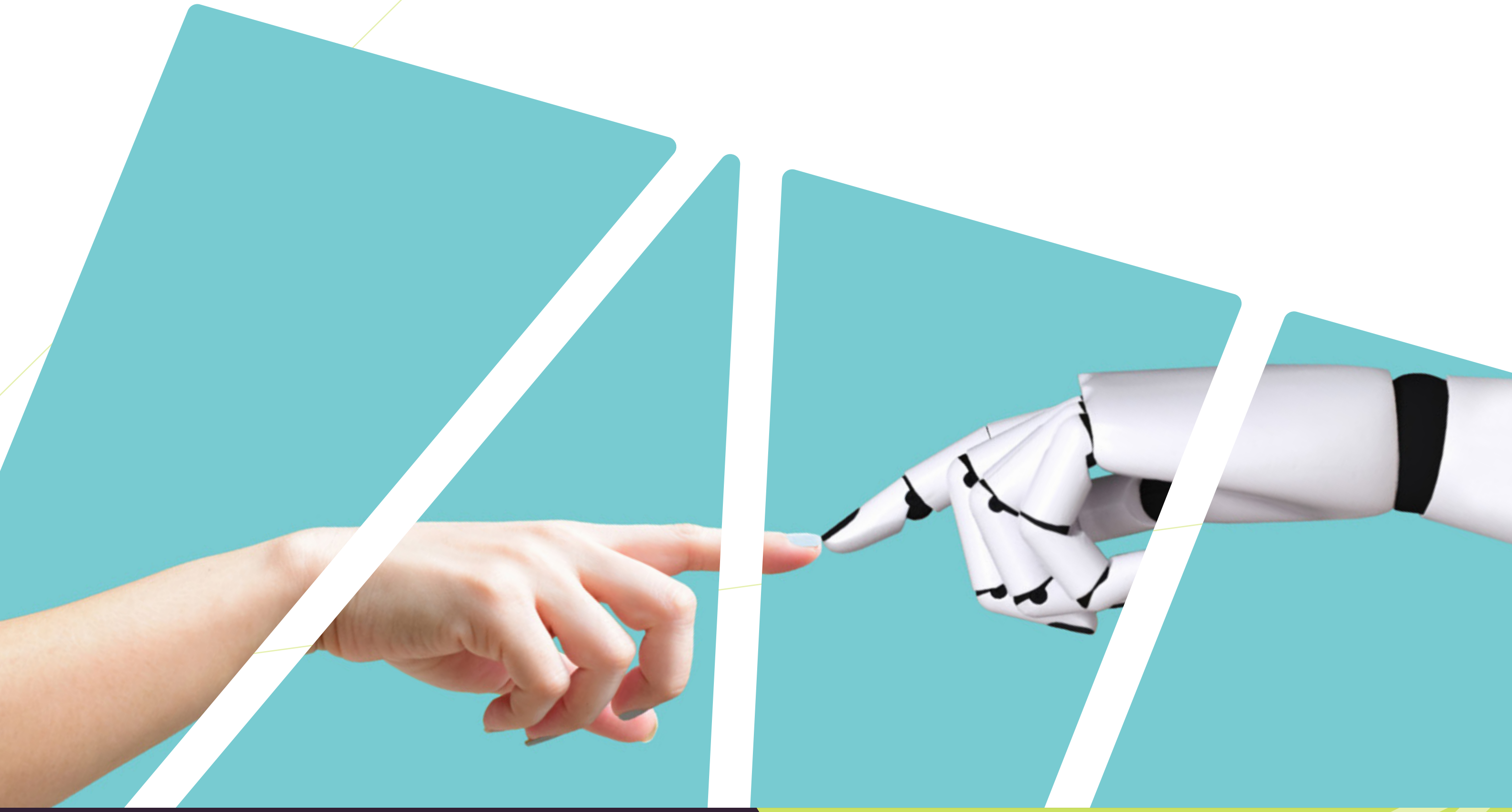
Or could it be the other way around? Is creativity actually data science being done on a much smaller, lower scale? If a machine can run different options, understand how people react to them and then move forward with that option. Isn't that what "creativity" means?

Indeed, some AI systems are now capable of creating very linear and repetitive stories with very standardised vocabulary and text structures — e.g. any news update on the stock market, weather, or sporting events — but anything more creative with fine subtleties triggering certain human emotions is still very much in the hands of humans, and to some extent will always be.

Data and AI can, however, play a very important supporting role to any translators, copywriters, creative strategists, etc.:

- Pulling data on the latest trends regarding a relevant topic
- Extracting relevant content segments from large databases for text analysis
- Conducting sentiment analysis on what people think via social media
- Monitoring large volumes of relevant videos and audio files to identify common cues
- Automatically creating relevant terminology and phraseology lists based on frequency analysis of language corpus databases
- Running completed text pieces against all types of filters to ensure that content is relevant, appropriate, and targets the right audiences

It is all about using data and AI in a smart way to supercharge your creative juices, alleviating the predictable and time-consuming repetitive tasks so that you can focus all of your energy upon what makes you human — idea creation and fine-tuning of your content to an individual style to reach a very specific group of people. It is about carefully balancing the amount of data and AI that you incorporate into your processes to fuel your production, but do not jeopardise creativity and the human touch.



Language in data science

Before moving on to a more specific summary of key data science processes used in multilingual content services, we need to explain some keywords which are essential when discussing language in the context of data science.

CL – 'Computational Linguistics'

Using computer science processes, models, and programming languages to understand text and react to any scientific questions concerning linguistics. There are complete MA degrees in this subject area that require a strong foundation in computer science and linguistics. It is an area that has grown massively, especially with the increasing importance of speech and language processing as Alexa, Google Home, and other smart speaker technology gain more and more popularity.

NLP – 'Natural Language Processing'

This is a key component of analysing language models (from basic frequency analysis to complete word models), creating training data with similar content structures and supporting the automatic creation of content. It is part of computer science and, more specifically, an essential area of artificial intelligence, as without it machines would neither manage to analyse language nor be able to produce anything meaningful.

SMT/NMT – 'Statistical Machine Translation/Neural Machine Translation'

SMT has been around for a long time, basically looking at a large volume of content, identifying frequencies and common themes, and then helping translators to produce translations automatically. Meanwhile, NMT is an evolution of the traditional SMT models, as it goes beyond merely looking at frequencies and simple 'what is before and after' word comparison models. Instead, NMT encodes each word into a numerical n-dimensional vector, with 'n' representing the number of connections that can be made between one word and other similar words (position, distance, frequency, etc.). This process is also called 'vector space modelling' or sometimes simply 'word embedding'.

For further information:
www.jigsawacademy.com/blogs/data-science/neural-machine-translation/

HPE – 'Human Post-Editing'

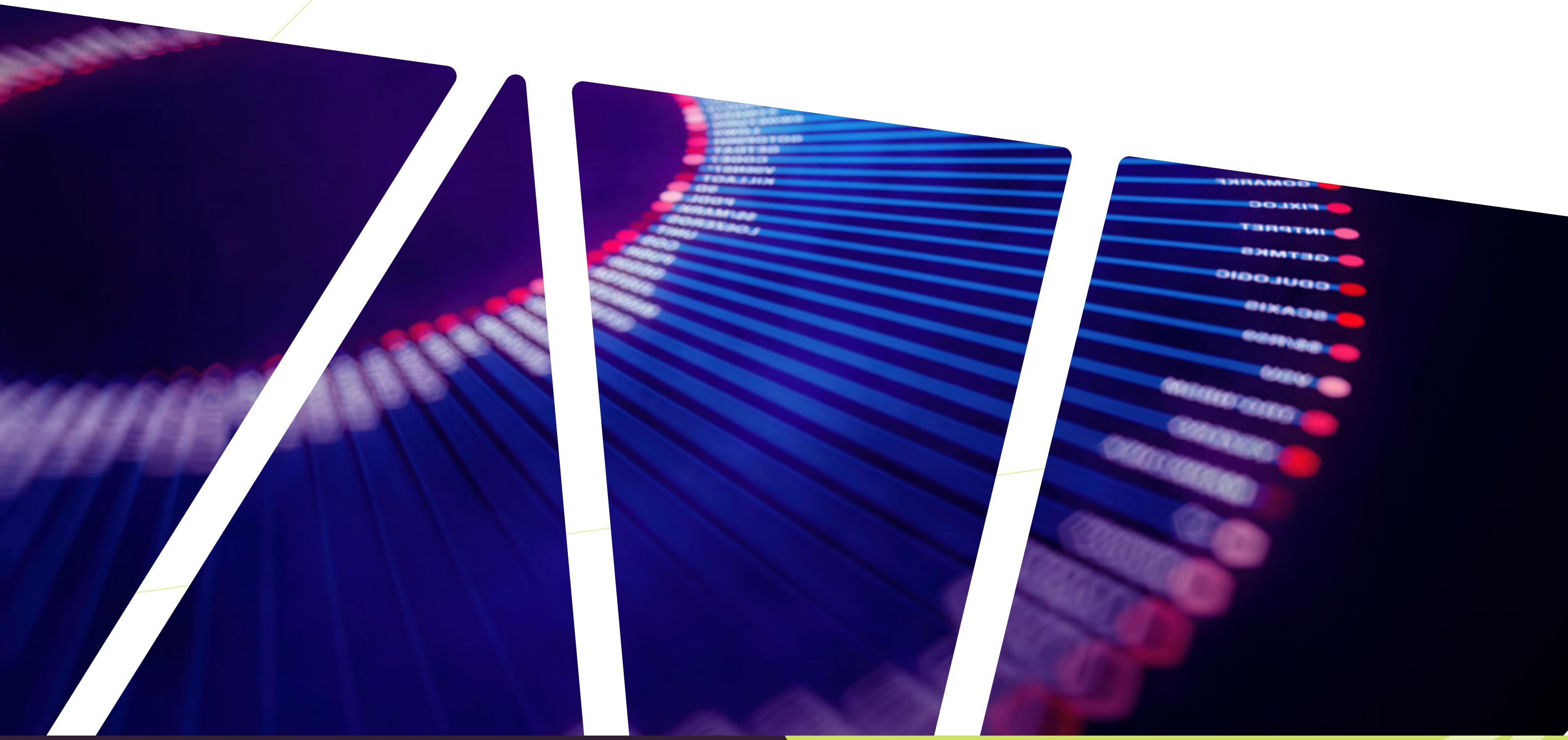
Taking any content that a machine has produced — be it via SMT/NMT engines or storytelling engines — correcting it, and fine-tuning it so that it is ready to be consumed by a human target audience, ideally without ever noticing that a machine has played an important role in creating the target content.

Python

There are (of course) many programming languages out there. Python is one amongst many, but it is the clear first choice for anybody working in computational linguistics, AI, and data science. It has a very extensive library of existing classes, functions and methods, which means that for many actions you only need to know the existing library term, import it, and run your data with it.

There are (of course) many more key tools, processes, technological solutions, and applications relevant to data science and linguistics, but this is not meant to be an exhaustive glossary of terms, but rather an overview of where data science is applied in multilingual digital marketing and content production.

On this note, let us move on to the core part of this article — practical applications of data science in language services, specifically multilingual research, production and optimisation in the digital content space.



Content filtering, creation and manipulation

In almost every desk job, we spend countless hours dealing with a huge number of files, often the same files repeatedly. Many of us do not recognise similarities and fail to repurpose previously completed files with similarities; instead, we start the same process each time that a new batch of files arrives.

Understanding the whole set of existing files in a relevant field, manipulating them to fit a new purpose, and then focusing only on the elements that actually need a refresh would lead to big gains in efficiency.

Automatic file filtering, creation and manipulation seem like very straightforward tasks which can be handled by Python or any other programming language. However, where machine learning and AI come in handy, is when you not only incorporate files done by yourself but also scrape the internet for similar pieces, extrapolate any elements of relevance, and combine them with your own work where possible. You could set up a number of rules with which to identify any relevant content, analyse its structure, and search for yet unknown ways of handling files that perhaps are much more advanced than what you have used so far. Thus, you would be able to stay ahead of the curve and present new and innovative ways of working to your teams and clients.

Another specific example would be in paid searches, where content analysts spend hours extracting keyword search query files regularly from the Google Ads interface, and then countless hours filtering the files by relevant/irrelevant terms, isolating each into distinct groups, and putting them back into campaigns where and when necessary. This is a massively time-consuming and arduous task. On top of that, if done in a rush by humans it is a highly error-prone and incomplete process.

As a first step, you could create simple word filters, which would allow you to cross-reference long lists of unwanted terms and delete those search entries. In a more complex operation, you could use NLP libraries in Python, gather all sorts of text — wanted and unwanted — and then create training data for each group (against which you can then check your search queries and categorise them accordingly). RapidFuzz, for instance, is an amazing Python library that can be used to compare two sentences extremely fast, even setting the ratio of comparison to make the algorithm more or less flexible.

Search queries come in all sorts of forms and lengths. Some are single words, while others are several words. Some make sense, while others do not. Some are grammatically correct, while others are not. In most languages, users will search for the shortest-possible way, meaning that for languages with many conjugations (how verbs change for different persons and tenses) and declensions (how nouns change in different positions of a sentence), words are cut down to their most basic form without any usually required grammatical modifications. Examining search activity in the top six languages in Europe over the last year in one of our biggest accounts with more than 650,000 organic keyword positions, we can see the following keyword length percentages:

1 word – 1.1%

2 words – 20.9%

3 words – 39.9%

4 words – 24.9%

5 words – 9.1%

Typical word combinations in the top six EU languages combine verbs in their infinitive form with adjectives and nouns. Every language is slightly different — German users, for instance, often omit articles or prepositions for simplicity. However, it is vital to focus on long-tail keywords to allow better targeting and less search competition. Those long-tail keywords look very different from grammatically correct phrases that you see in properly written text:

Sneaker	135,000 SV
Der Sneaker	20 SV
Sneaker schwarz	5,400 SV
Sneaker schwarz Damen	9,900 SV
Sneaker kaufen	2,400 SV
Sneaker kaufen online	720 SV

When optimising your search account, a linguist could go through the many thousands of search terms that a search query report (SQR) will produce, filter what looks relevant, and add it to the account or (if unwanted) ensure that it sits in a so-called negative list, which means that keywords with those terms will not trigger adverts anymore. Undertaking this process across many languages daily is expensive and inefficient.

So, how can automation help here? You could use NLP to take all possible variations of each search term with the help of a k-skip-n-gram algorithm, which sounds complicated but really is not so difficult to understand. N-gram refers to the maximum number of consecutive word combinations possible within one search term.

Example search query:
buy red sneakers online

You have four possible gram structures:

Unigram	Bigram	Trigram	Quadrigram
buy, red, sneakers, online	*buy red, red sneakers, sneakers online*	*buy red sneakers, red sneakers online*	*buy red sneakers online*

That is a total of 10 split words/word combinations. But there is a problem here: what if we want to take non-consecutive variations into account as well?

After all, “buy sneakers” without “red” is very much a valid search term. In such cases, you can resort to the k-skip-n-gram model, which allows you to skip a k-number of terms.

So, if we extend the aforementioned example with k-skip-n-gram variations, we additionally obtain:

buy sneakers, buy online, red online

One step further into this analysis would be to scramble the words in the Unigram and therefore get all the possible outcomes in the Bigram, Trigram and Quadrigram. Comparing the results against each other means that no stone will remain unturned.

As a next step, the analyst would filter out any unwanted variations through the use of pre-existing filters and then cross-reference the output with all existing keywords in the account. Whatever is identified as new can then be picked up by a linguist and added to where it best fits.

Search advertising constantly evolves and requires you to quickly adapt to any new tools, changes in advertisement regulations, and updated requirements regarding keywords and advert copy creation. The objective for the content analyst is to focus their energy on the creative elements and the fine-tuning of content to the expectations of their clients — be it for reasons of style or performance requirements.

Language filters using training data

We have briefly spoken about simple filters and training data in the previous section. Machine learning and all forms of NLP are heavily dependent on massive amounts of content to analyse to produce relevant output. Machine-produced content could be entirely originated, translated, or use a hybrid approach in which text elements created by humans are merged and completed by a machine, or vice versa (i.e. a human fine-tunes computer-generated text segments into a cohesive target piece).

The most important first step before any machine-generated target text creation is to collect and then clean the data. This may sound simple but is, in fact, a highly complex and onerous process. Searching for and deleting any pejorative or other unwanted words is easy as long as your list is extensive enough. You could expand your source list with synonyms or other related terms by cross-referencing your source terms with an online thesaurus.

Before you can play your filter list against all of your gathered content, you need to ensure that all words in the filter are trimmed down to their stem and/or lemma. Those are basic processes that are essential during content cleaning. The following example allows visualising the difference between stemming and lemmatisation.

If you use the popular PorterStemmer stemming tool in the NLTK library you will obtain the following output:

Languages › **languag**

Connections › **connect**

'Languag' is not a real word but could be used as a filter. It would, however, be better to reduce words to their dictionary form, which is where a lemmatiser comes into play:

Languages › **language**

Connections › **connection**

In short, stemming and lemmatisation are ways in which to clean any textual data and ensure that you have identifiable and searchable core terms in your final list.

There are many more steps in data science and, more specifically, NLP to prepare your content for analysis and production. One such step is tokenisation, which splits text into words or sentences depending on the requirements and purpose. You also have specific Python libraries for the removal of stop words, which refer to words with a very high frequency and little meaning, such as articles ('a', 'the') and conjunctions ('and', 'or', 'for', 'but', etc.). A great source of additional information on processing tests is that of data science websites such as Towards Data Science[3].

After cleaning your source content data, you need to convert all of the words into something that the computer understands; in other words, they need to become a number. However, one simple number per word would not be representative of the complexities that a word expresses — you have the meaning of the word standing not only by itself but also in combination with other words directly before and after as well as in relation to any other words in any given piece of content.

This process of assigning a number to a word is called word vectorisation, word embedding or vector space modelling. It is the same process as explained above when discussing statistical/neural machine translation, as this is a crucial step in optimising your machine-translated target content. Python has many existing libraries to undertake all of the hard work for you. It basically automatically assigns a number to every single element relevant to a given word — be it frequency, position, order, meaning, etc.

The combination of all the numbers — each representing one dimension — ultimately results in a multidimensional, unique vector describing a word in a given context to a machine.

The question now is concerned with how pre-processed, vectorised text helps with filtering. The answer is simple: let us say that you have collected a large database of content representing typical spam and want to use this content pool and cross-reference it with any incoming email to check whether it is spam. To do so, you need to first process and vectorise your pool and convert it into training datasets. You can then use Python to cross-reference any of your incoming email against that training data; if similarities occur, you can flag it as spam.

You could also gather all content written by a certain author and/or in a certain style, convert it into a training dataset, and then play against new content to identify similarities. This process not only can be used to filter out unwanted or desired material but also is extremely useful in grouping content into different categories.

It is important to note here that all of those automated text processing methods are by no means perfect. Remember that the explained machine learning models use statistical modelling to identify how relevant content is to a requested task. The greater the amount of content in your training data, the better it is, which often means that English is much better built out than any other language.

Even the biggest corporations in the world with unlimited funds cannot yet fully tackle the complexities of identifying specific speech or content types across many languages. Just recently it has been revealed by a senior Facebook employee that Facebook has thus far failed to filter out huge amounts of hate speech and misinformation[4].

A human-in-the-loop (HILP) process combining machine learning methodologies to pre-process and filter with human input throughout is, ultimately, the only way in which to achieve almost 100% content accuracy.

Where else are some of the NLP processes explained above useful in marketing?

Sentiment analysis

In PR, communications, marketing and advertising, we constantly search for the best ways in which to track, monitor and understand how customers perceive products or services and, consequently, how we can improve the areas of greatest interest.

Only a decade ago, anybody in PR who wanted to analyse multilingual content pieces across different publications needed a team well versed in MySQL and/or Microsoft Access to gather as much content as possible and structure it in one centralised database. One of the challenges was that a large proportion of the content was from offline magazines, newspapers, radio interviews, etc., which meant that analysts had to manually sift through all of the material and insert key components in relation to when, who, in what context and how somebody described a brand's product or service. Once the manually 100%-human-led analyses had been completed, you would use the database to create models and graphs that explain trends. As you can imagine, it required a huge amount of time and concentration from analysts who would work tirelessly across several languages.

Regardless of how meticulously everybody worked, mistakes inevitably happened and the final data was arguably at least partially imperfect.

Luckily, we have experienced quite a revolution over the last 10 years in terms of the digitisation of media, with most media now being online or at least offering both. Aside from the obvious environmental benefits, the collection and analysis of such data have become considerably easier.

But how about the analysis of sentiment, which is ultimately the key part that you want to achieve when analysing feedback, comments and reviews?

Thankfully, Python has also produced some fantastic libraries (such as NLTK and TextBlob) which support sentiment analysis functions as well as the usual data cleaning and linguistic processing tools explained in the previous chapter.

Why is sentiment analysis so difficult for a machine?

There are obvious words in each language that indicate positivity, negativity or neutrality. Using some basic NLTK tools combined with a broad corpus would easily allow you to run classification across the three categories.

It becomes more difficult when working with sentences:

For example: *"It is not really that great."*

"Great" is easily classified, but the negation "not" makes it more negative. The adverb "really" softens the negativity and conveys a level of uncertainty but does not quite neutralise the statement. As you can see, it is not so straightforward.

Users often will use several sentences to express their views, in which case, several components of the whole statement have to be factored in to identify the right sentiment:

For example: *"At first, I was really confused because the story seemed complex and hard to follow, but as I kept watching, I developed*

an interest and liking for the characters, their interactions with each other and the great things they achieved."

Any human would classify that as positive, but this sentiment is not easy to grasp for a machine. "Confused" could be identified as negative, but "complex" and "hard" do not necessarily mean that you dislike something. You may love complex and hard-to-solve puzzles, in which case it is a strong positive sentiment.

In such difficult cases, the machine needs to look at the frequency and relationship of different words with one another and compare them to existing training data. The naive Bayes classifier — a popular statistical classification method integrated into the scikit-learn library in Python — takes a set of pre-classified example sentences as training data and then calculates the probability of any new piece of content belonging to any of the pre-classified training data classes.

As you might guess, this is not 100% perfect, but it has become an excellent way in which to categorise large volumes of content pieces by sentiment, such as in social media, brand monitoring, PR analysis, or market research.

The larger and more detailed the dataset for your training data is, the better it is, and there are some massive and amazingly rich databases available which continue to grow daily[5].

Sarcasm, tone, non-speech acts such as gestures and body language, and comparatives ("Oh, this laptop is much thinner than yours" is not necessarily negative and, therefore, is not clear to the machine) all constitute a challenge to automatic sentiment classification, but at least in English an increasing amount of impressively extensive libraries and corpus data manage to achieve outstandingly accurate results.

As with many other developments in NLP, aside from English, all of the aforementioned processes lag behind in other languages due to a limited amount of data.

You could (of course) use NMT to translate all foreign-language content into English and then use one of the popular processes well established in English for analysis. This process, however, may lead to two possible reductions in accuracy — firstly, due to the translation and, secondly, due to the accuracy of the sentiment analyser itself.

Data collection and annotation

Since we have spoken a lot about training data, let us explore why it is so important and how much of the collection and classification of data can truly be automated as opposed to having to be done by humans to guarantee reliable results.

For training data to be truly automated, it is extremely important to test it before making a prediction, to escape the common error of overfitting. The intention is to create an algorithm that could actually predict the future and not just explain the past.

Let's use an example to explain this concept. If we are teaching a self-driving vehicle to drive with information showing a turn sign and at the same time, a bird passed by the sensor, it would be a terrible performing algorithm if it understood that each time a bird passes by, the vehicle should make a turn.

If we test our training data, we will identify these abnormal situations within our dataset, providing value on how well it has been trained and therefore how well it will perform.

The global data annotation market has been growing massively over the years and is expected to reach USD 3.4 billion by 2028, with an estimated compound growth rate of 27% between 2021 and 2028[6].

Such growth is driven by an insatiable demand for development in autonomous driving, smart speakers, chatbots, face recognition, automated translation and interpreting services, robotics, etc.

AI and its machine learning models can only produce reliable results if the training data is large enough and accurately annotated. Data can come in all sorts of formats — text, speech, video or image — all of which must be labelled so that the machine can understand its meaning and use it for analysis.

For text, there are already some pretty advanced OCR ('optical character recognition') technologies available. Most of them are heavily geared towards the Latin alphabet, whereas there is still a long way to go in terms of accuracy for more complex alphabets or characters.

Speech recognition has received a massive boost in research and development over the last five years, thanks to growth in the popularity of smart speakers and their built-in voice assistants (including Amazon Echo, Google Home, Apple's Siri, Baidu Xiaodu Zaijia powered by DuerOS, Yandex Station with Alice, Alibaba, and Naver ClovaAI, just to name some of the biggest names in the market right now).

And there is no end in sight for further growth and optimisation of AI-enabled speech assistants.

Just over the last year during the coronavirus pandemic (2020/21), smart speakers and smart display sales have increased by 34.8%, reaching 39.5 million units[7].

One question that many ask is where smart speakers obtain their information. For general questions, information is pulled from whichever search engine is the most popular in the respective market, or from platforms such as Wikipedia. For any information on businesses and addresses, Google Home will retrieve its information from Google My Business pages, or from websites like Yelp in the case of Amazon Echo or Apple's Siri.

If you are not happy with any of the results provided by smart speakers or displays, you can optimise results, although the process of doing so is still very manual:

1. Identify long-term keywords, mostly phrases, and sentences relevant to the respective client. Remember that keyword structures are very different in voice search from in general search engine queries. Agencies like Locaria have tested numerous variations in voice search to understand what syntax is most likely to obtain relevant results.

2. Have someone test all of those keywords with the smart speaker/display, note down the results, and identify gaps, incorrect results, or competitor responses.

3. Develop a roadmap with near-term recommendations. It is important to react quickly, and results can change quickly. Recommendations will differ depending on the smart speaker, but mostly they are a combination of content updates on the brand website, Wikipedia, or any third-party business pages such as Yelp, Yext, or Google My Business.

4. Test and learn. Measuring content effectiveness in voice search is still tricky, as it is difficult to tie back improvements in traffic or ROI to voice search optimisation strategies. The best way is to repeat the manual auditing process a few times per year to understand how results have improved. Also, monitor traffic improvement in on-site content specifically optimised for voice search.

Now, how do Amazon, Google, Apple, Baidu, etc. improve their speech recognition systems? While machines have become much better, human reviewers are still a key part of the process of continuously enhancing speech recognition training data and NLP systems. All of the big players admit to using humans in the loop to ensure accurate and relevant results[8]. Those reviewers not only filter out unwanted material but also are responsible for annotating voice clips to classify speech segments more precisely. PII (personable identifiable information) is not provided to the reviewers; instead, IDs are assigned to users and voices are distorted.

The human-in-the-loop process is very common across all aspects of data annotation. The objective of technology companies is to automate as much as possible while having humans monitor the process throughout, rectify inaccurate segments, and fill in gaps. More information on data annotation and how machines and human processes are woven together to achieve better machine learning results can be found in the following articles:

https://medium.com/anolytics/what-is-data-annotation-and-what-are-its-advantages-95766213351e

https://www.bmc.com/blogs/data-annotation/

Incrementality testing

One of the major challenges in multilingual content production has been that output is largely judged based on subjective criteria. Those criteria may be client-specific, sometimes based on a few customer reviews or even on an individual or small group of linguists. None of those cases represents the true success or failure of the end content piece; rather, they constitute isolated views. It requires large volumes of data from different sources monitored and compared over a significant amount of time to truly identify statistically significant results.

If we want to be strict and fair about the content that we create for our target audiences, we need to put aside personal views and emotions, which may even have been built over many years of working in a relevant specialised field, because, ultimately, they are still not necessarily an objective reflection of what the end-users feel when reading your personally created target piece.

How can such challenges be overcome? Are there ways in which to set up testing that provides a true understanding of the impact of content on end-users?

This is where Performance Linguistics® and data science come into play.

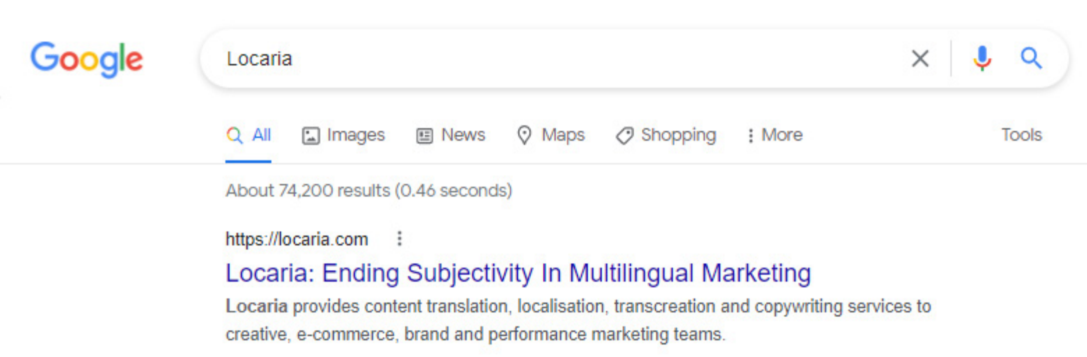
Performance in multilingual content production is not channel-exclusive; paid media localisation will impact organic media; website localisation is directly linked to SEO activity; and CRM translations will drive open rates and, ultimately, website traffic.

It is a continuous cycle demanding digital marketing tests, with reporting woven into the many different ways of creating content for multiple markets and languages. Having cross-channel visibility enables specialised agencies (such as Locaria) who have mastered the combination of language expertise, creativity, and data analytics to make better recommendations and determine important strategic decisions for brands worldwide.

Building on a data-led Performance Linguistics proposition, it has been possible to further evolve and truly meet the increased customer demand, requiring not only expertise in language services but also being able to use exceptional digital and data expertise to deliver multilingual content that drives value. With data, agencies can prove the value of language services and through shared tools and resources, they can gather data on digital performance to innovate language solutions.

One area in which incrementality analysis can truly help to understand performance is the localisation of SEO meta content. In the following example, we are looking specifically at meta title tag content structures. The project aimed at exploring how much a page's title tag could influence SEO traffic. Page titles are extracted to the SERP, where they are presented in bold, blue, and underlined text. They are the showcase of a page and are entirely malleable.

The objective of the project was to determine the optimal page title syntax for any given product page within the client's catalogue.



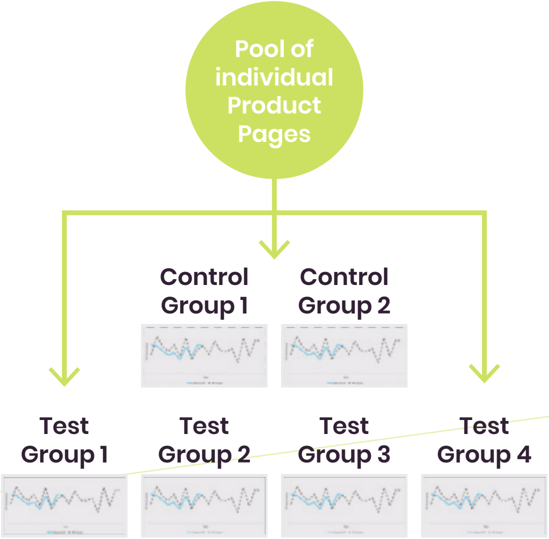
About 74,200 results (0.46 seconds)

https://locaria.com

Locaria: Ending Subjectivity In Multilingual Marketing

Locaria provides content translation, localisation, transcreation and copywriting services to creative, e-commerce, brand and performance marketing teams.

The data science team wanted to compare how four different syntaxes performed in driving traffic. Each syntax was structured around a unique permutation of a product's attributes (colour, sport, sub-brand, etc.). Thereafter, each syntax would be allotted to one of four product page groupings. The groups were carefully formed to be homogenous, meaning that they needed to contain the same number of products, a similar balance of gendered and unisex products, and drive comparable levels of traffic. The challenge lay in creating a design in which we could minimise most variance between groups, ensuring that any difference in their performance was due to the title change.



Groups had to be built around homogenous traits, but we wanted them to express divergent performances after the title change, which meant that the chosen syntaxes had to be different enough from one another to drive different behaviours. The analysts worked closely with SEO-trained linguists to select a combination of short vs. long titles and product-specific vs. brand-specific attributes, as well as different permutations of these. All groups mentioned the product name.

Examples of varying title tag syntax would be:

product name + Sub Brand | Brand

category name + Product name + gender | Brand

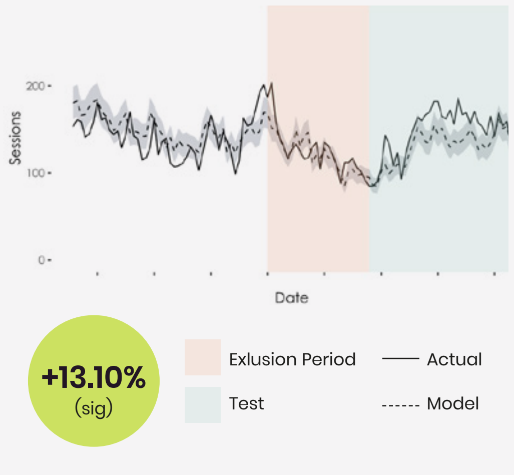
product name + category name + sub-brand | Brand

Same structures but with a brand at the beginning are also an option.

Each test group needs to have a control group to understand what actually will happen in terms of performance in existing structures that remain untouched.

Based on past performance and trends, a hypothesis regarding the test groups needs to be forecasted against which the actual performance following the test can be compared to identify the true incremental impact.

A typical results graph from a test group would appear as follows:



After observing the results over several weeks, it was discovered that two syntaxes successfully drove in significantly more traffic than did the others, with one clear winner. The top-performing syntax gathered 13% (with 99.4% significance) more sessions than would the old syntax. Two other groups had a stable performance, attesting that any title change was not sufficient to generate a change in sessions either upwards or downwards.

Engineering a statistically sound framework designed and tested by a team of data scientists, SEO specialists, and linguists allowed the team to make an informed decision to identify and follow the optimal title for product detail pages. This strategy guided linguistic choices by leveraging data science and analysis.

This is only one way of using incrementality testing to prove the true impact of your content. You can use a set of tests structured smartly and strategically via a long-term roadmap to create complete cross-channel measurement frameworks which allow you to not only understand DTC last-click digital performance but also prove the performance incrementality brought about by branding, creative, possibly offline, or other localisation activity.

What is the impact of the switching of certain paid search campaigns upon organic search? Do traffic and revenue decrease? Or do your organic results capture all of the traffic?

Does your branding display or organic social activity truly drive incremental revenue? How much of your ROI can be directly attributed back to those channels which are not always fully tracked from engagement to conversion?

The aforementioned advanced testing framework and strategies are the solution to all of those questions.

Conclusion

It has been an exciting few years in the language service industry space, particularly for those agencies specialising in digital content. In most parts of the world, the majority of content is already online, and over the last couple of years, even brands that have loathed the idea of giving up on traditional media formats (such as magazines, billboards, and linear TV) have started to shift the majority of their budget to online channels.

At the same time, technology in the language production space has advanced in leaps and bounds. No one can deny the progress that machine translation has made, especially with the rise in neural machine translation technology (which grows its training datasets across multiple language pairs by the second and, consequently, produces ever more impressive results).

Will it ever surpass the human mind? The need for a final human touch, providing creative input and direction and fine-tuning the final product for specific audiences, will not simply fizzle out. Our human minds are difficult to track and often change without logic in totally unexpected ways. We need to learn to use technology for our benefit and embrace it wherever and whenever we can to be able to manage the vast amount of data that we have to process, analyse and repurpose in our daily life.

There are (of course) many concerns: English has received much attention from the major players in technology; thus, any natural language processing involving English is still much more built out than any other language. The major European and Asian languages produce great results in neural machine translation when translated from or into English and are improving daily, but there are a plethora of languages that lag far behind and extensive research and work will be required for them to catch up. Not every language will receive funding and attention from big players, because, ultimately, much depends on how much revenue those languages generate. Thankfully, the digital world manages to connect us to almost all parts of the world, even the smallest communities and languages, and through that valuable information on many cultures and languages is stored online and can be picked up for analysis whenever we need access to it.

Measuring the success of content, especially as human, machine and hybrid-model solutions start to blend together, will become even more important. We cannot rely on the views of a few or on the subjective opinion of a small group of linguists.

Embracing the many tools and processes that data science offers us will help to remove subjectivity from the multilingual content creation process and provide truly objective reports on what works and what does not.

There is (of course) risk when using large datasets in data science to identify trends and behaviours and then make possibly long-term decisions. Analysts must be trained not only in statistics, coding and analysis but also in how to carefully select accurate data and ensure that it is relevant and fairly distributed.

What if, as a consequence of a test, certain races receive different treatment from others? Those kinds of ethical considerations need to be made and analysts must be educated on how to carefully integrate checks throughout all steps of their workflow[9].

The bias bounty challenge report from Twitter in 2021 proved that even the biggest technology companies still have much work to do and cannot stop for their systems to continuously evolve to guarantee fair results[10].

The bias bounty challenge has revealed several issues:

- An image cropping feature removed Black faces in favour of white faces
- Models favoured encoded stereotypical beauty standards — slimmer, younger, feminine, and lighter-skinned faces
- Linguistic biases between how Twitter handles English memes and Arabic script memes
- Twitter's model showed a preference for emojis with lighter skin

One company alone will never be able to tackle all of those challenges in a centralised way. It will require the concerted effort of a wider ethical AI community to constantly feed into the process and optimise algorithms further.

There is one thing that we always have to remember when we become excited about the use of any technology involving data science and natural language processing:

Machine translation, automated copywriting, sentiment analysis, and so on — all of those things work well when they are based on large volumes of constantly expanding and optimised data — data produced professionally by humans! Without human input, the kind of accuracy in machine translation, data annotation, machine learning, and natural language processing that we see nowadays would not be possible, simply because language changes all of the time and does so at speed and often unexpectedly without any warning.

It is the job of specialised agencies and companies, experienced linguists, content analysts, experts in performance linguistics, and data scientists with a passion for languages to strike a balance between machine-driven and human-driven processes to provide end clients with content that they trust, enjoy and cannot wait to see or hear more of.

References

Page 4 1. https://towardsdatascience.com/machine-translation-b0f0dbcef47c

Page 5 2. https://slator.com/new-generation-of-data-scientists-tackles-translation/

Page 13 3. https://towardsdatascience.com/nlp-text-preprocessing-steps-tools-and-examples-94c91ce5d30

Page 14 4. https://edition.cnn.com/2021/10/26/tech/facebook-papers-language-hate-speech-international/index.html

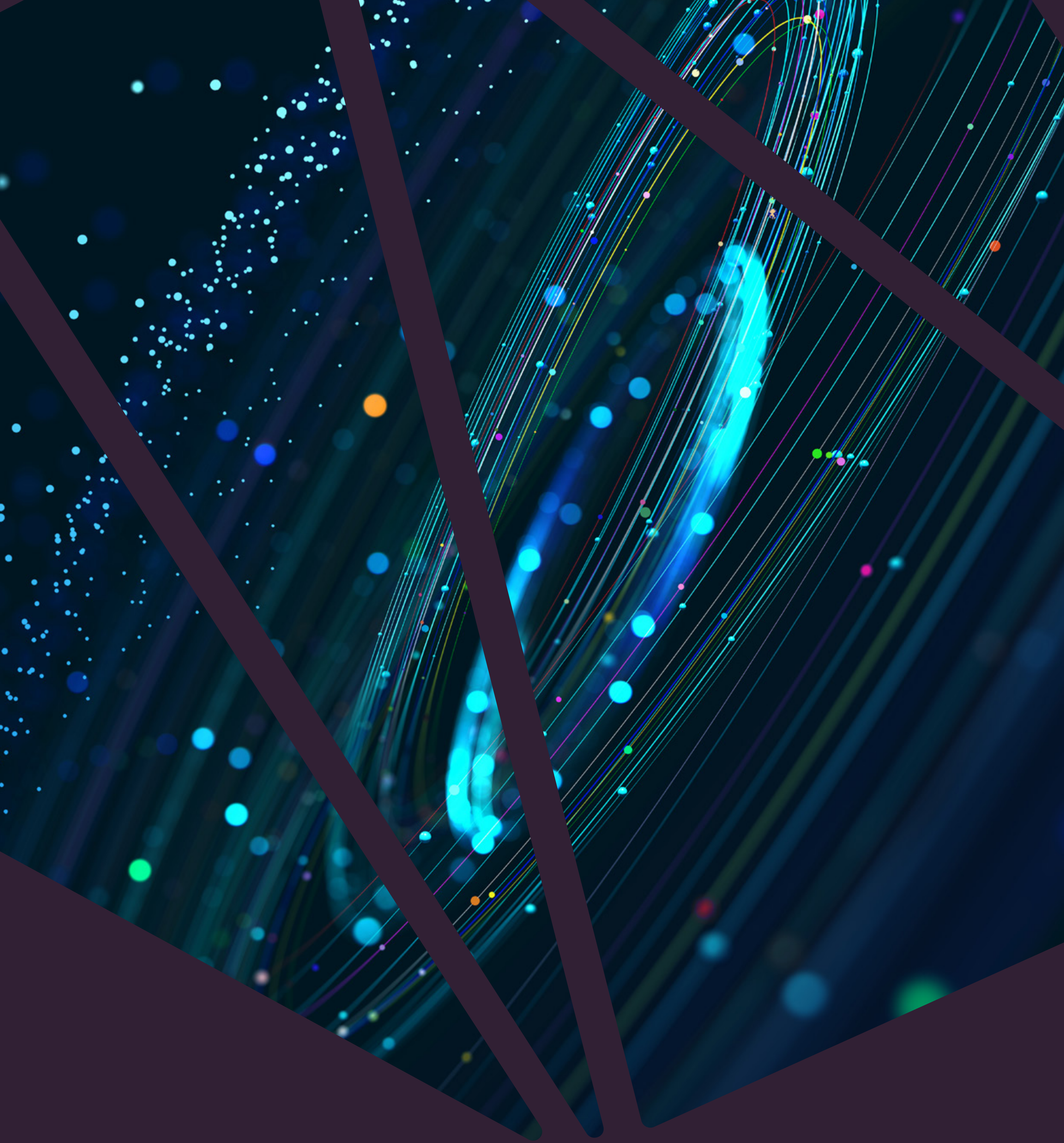
Page 16 5. https://analyticsindiamag.com/10-popular-datasets-for-sentiment-analysis/

Page 17 6. https://www.businesswire.com/news/home/20210917005169/en/Global-Data-Annotation-Tools-Market-and-Segment-Forecasts-2021-2028-A-USD-3.4-Billion-by-2028-with-CAGR-of-27-Forecast-During-2021-to-2028---ResearchAndMarkets.com

7. https://www.businesswire.com/news/home/20210914005799/en/Strategy-Analytics-The-Smart-Speaker-Markets-Recovery-is-in-Full-Swing-as-Shipments-in-2Q21-Surged-to-Record-Levels

Page 18 8. https://www.bbc.co.uk/news/technology-47893082

Page 24 9. https://www.nextgov.com/ideas/2021/09/data-science-education-lacks-much-needed-focus-ethics/185204/ 10. https://www.zdnet.com/article/twitter-algorithmic-bias-bounty-challenge-unveils-age-language-and-skin-tone-issues/



Locaria

www.locaria.com