DZone
A DEVADA MEDIA PROPERTY

DZone > AI Zone > How to Build Your Own Text Classification Model Without Any Training Data

How to Build Your Own Text Classification Model Without Any Training Data

by **Shashank Gupta** 👤 MVB · **Jan. 25, 18 · AI Zone · Tutorial**

One of machine learning's most successful paradigms is supervised learning, which lets you build a generalization model by learning from a lot of training examples. Supervised learning is extensively used in natural language processing to build multi-class or multi-label text classifiers for solving a variety of use cases like spam detection, sentiment analysis, emotion analysis, customer intent analysis, etc. Anyone familiar with the process of building a text classification model would know that it involves three steps listed below:

1. **Prepare training data**: Training data, also known as labeled data, is a corpus of domain-specific cases that are manually annotated with labels that the classifier is expected to classify.

2. **Training the text classification model**: A suitable machine learning algorithm is then selected to train a model on the labeled dataset. We have covered some of the techniques in our previous blog for building a sentiment analysis model.

3. **Testing and validation**: A portion of the labeled dataset is set aside for testing and validating the trained model to assess its generalization ability.

The traditional supervised learning paradigm breaks down when we do not have sufficient labeled data for our task to train a reliable model. Text classification projects in the real world often face their biggest hurdle in getting access to a sizeable amount of data and manually annotating them. As a result, most real-world projects get stuck at the first step. Those that *do* cross the first hurdle often find their data to be either insufficient or skewed toward certain classes, leading to biases in the resulting AI models.

What Is Custom Classifier?

A lot of research in deep learning is directed toward learning better representations from less data, and one such exciting field is zero-shot learning. Quoting Ian Goodfellow from

his Quora answer, "zero-shot learning is being able to solve a task despite not having received any training examples of that task."

In other words, imagine predicting the emotion of a tweet without providing any training examples of tweets with that emotion label. ParallelDots' in-house research team came up with its own paradigm of zero-shot learning for text classification and released a research paper on the same, aptly titled Train Once, Test Anywhere: Zero-Shot Learning for Text Classification.

Additionally, living up to our commitment for bringing AI to your fingertips, we decided to move this exciting piece of research into a commercial application and launched a new API that lets you utilize the full power of zero-shot learning for your text classification use-cases. We are calling it Custom Classifier since it lets you build your text classifier on your defined, custom categories. This is a revolutionary step in our pursuit of developing cutting-edge AI solutions to let users develop and integrate custom text classification models without the needing to build any training data.

Learn by Example: Build a Sample Text Classification Model

Since no training data is required, this brings down the cost and time it takes to build a text classification model. With all the background knowledge in place, let's try to see the custom classifier in action.

In the below screenshot taken from the Custom Classifier demo page on our website, we took a sports headline and defined five categories that we want to classify this headline into: World Politics, Sports, Religion, Entertainment, or Business.

TRY OUR CUSTOM CLASSIFIER DEMO

ENTER A TEXT

Thrilling contest between Manchester City and Manchester United ends in a draw.

ADD YOUR CATEGORIES

world politics x   sports x   religion x   entertainment x   business x   Press Enter to add text   ANALYSE

RESULT

SPORTS   16.32 %

WORLD POLITICS   5.68 %

BUSINESS   1.11 %

ENTERTAINMENT   0.35 %

| RELIGION | 0.19 % | |

As you can see from the results, the Sports category has the highest probability score of all and therefore is the most likely category of our input text. Please be mindful of the fact that we defined the categories at the time of testing and the underlying AI model was not explicitly trained on them.

Next, we try to further break down the sports category into different categories like soccer, golf, etc. and see if our classifier can understand them.

ENTER A TEXT

Thrilling contest between Manchester City and Manchester United ends in a draw.

ADD YOUR CATEGORIES

world politics x   sports x   religion x   entertainment x   business x   soccer x   golf x

tennis x   Press Enter to add text

ANALYSE

RESULT

SOCCER	48.85 %	
SPORTS	16.32 %	
GOLF	7.81 %	
TENNIS	6.29 %	
WORLD POLITICS	5.68 %	
BUSINESS	1.11 %	
ENTERTAINMENT	0.35 %	

Again, we see that the classifier accurately identifies the Sports categories without being explicitly trained on them. I hope that by now, you have started to understand the potential of this technology.

As the last step, let's try to further break down our Soccer category into different types of leagues that are played around the globe and see if our classifier can correctly predict the same.

TRY OUR CUSTOM CLASSIFIER DEMO

ENTER A TEXT

Thrilling contest between Manchester City and Manchester United ends in a draw.

ADD YOUR CATEGORIES

world politics x   sports x   religion x   entertainment x   business x   soccer x

premier league x   bundesliga x   golf x   tennis x   Press Enter to add text

ANALYSE

RESULT

PREMIER LEAGUE	96.17 %	██████████████████
BUNDESLIGA	61.27 %	████████████
SOCCER	48.85 %	██████████
SPORTS	16.32 %	███
GOLF	7.81 %	█
TENNIS	6.29 %	█

Magical, indeed! Our classifier correctly picked up the league, as well.

In summary, Custom Classifier gives you a glimpse into the future of text classification where very few or no training examples will be required to classify a piece of text into custom defined categories reliably. Such an ability will open up endless possibilities and give a new life to numerous dormant text analytics projects that could never have seen the light of day due to lack of training data or training resources.

Set Up a Custom Classifier

Setting up a custom classifier is very easy and can be done in three easy steps:

1. Sign up for a free ParallelDots API account and log in to your dashboard.

2. Navigate to the custom classifier section in your dashboard, provide sample text, and define some categories to analyze your text. By default, you will be in test mode, which lets you tweak your category list as many times as you want for no cost.

3. Once you are satisfied with the classification accuracy, click on **Publish** to deploy your classifier for production use and get classifier ID. Once published, you will be able to access your Custom Classifier via the API endpoints.

API USAGE	API KEY	CUSTOM CLASSIFIER	EXCEL ADD-IN	PAYMENT METHOD	BILLING HISTORY

TEST MODE

> ❶ Test Mode allows to modify the categories to optimise them according to your data. Please note that to achieve best classification results, provide category list which is mutually exclusive. Check out our FAQs below.

ENTER CATEGORY NAME:

Sports News

ENTER TEXT:

Thrilling contest between Manchester City and Manchester United ends in a draw.
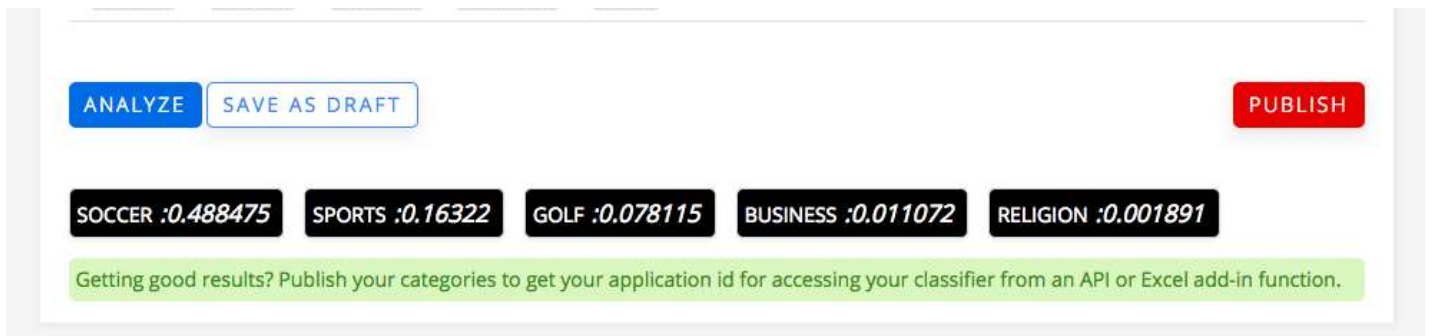
ENTER CATEGORY:

sports x	soccer x	religion x	business x	golf x

API documentation for the Custom Classifier is available here. An Excel add-in function will be available soon to use it from MS Excel.

In this post, we introduced a new paradigm in text classification, and we hope that our users will benefit from it tremendously. We are already seeing some exciting use cases being solved by Custom Classifier by our users, like classifying job postings, categorizing news articles, processing open-ended survey feedback, etc. Often, the very first step to analyze raw data is to categorize it into categories that you care about and monitor trends over time to gain actionable insights.

Advantages of Custom Classification Over Standard Text Classification

Standard text classification models like sentiment analysis are not always sufficient to get a holistic view when analyzing unstructured data like the voice of the customer. Custom Classifier, on the other hand, can give you the ability to categorize such data into more nuanced categories that we care about. For example, using Custom Classifier, a fast food chain owner can categorize the feedback received from customers via social media into categories like Price, Ambiance, Staff Behavior, Food Quality, etc. to gain better insights into what aspects of her business she needs to improve. Using a simple sentiment analysis, she may know how her business is doing overall but could potentially miss some critical trends such as how staff behavior and the footfall rate correlate. The potential of Custom Classifier is to classify text beyond sentiment labels (positive, negative or neutral) so that you not only know *how much* negative feedback you got on your product but also what users *talked about* in those negative statements.

We believe Custom Classifier can empower users with no background in data science to build and deploy their own classifiers and gain an edge over their competitors. It will also help enterprises radically transform themselves into data-driven organizations and will ultimately accelerate the industrialization of artificial intelligence and machine learning. You can also explore more text classification blogs here.

Like This Article? Read More From DZone

DZone Article
Simple Sentiment Analysis With NLP

DZone Article
Custom Text Classification in SmartReader

DZone Article
AI vs. Machine Learning vs. Deep Learning

Free DZone Refcard
Introduction to TensorFlow

Topics: AI, ALGORITHM, DEEP LEARNING, MACHINE LEARNING, NLP, SENTIMENT ANALYSIS, SUPERVISED LEARNING, TEXT CLASSIFICATION, TUTORIAL

Published at DZone with permission of Shashank Gupta , DZone MVB. See the original article here. ↗
Opinions expressed by DZone contributors are their own.

ABOUT US
About DZone
Send feedback
Careers

ADVERTISE
Developer Marketing Blog
Advertise with DZone
+1 (919) 238-7100

CONTRIBUTE ON DZONE
MVB Program
Zone Leader Program
Become a Contributor
Visit the Writers' Zone

LEGAL
Terms of Service
Privacy Policy

CONTACT US
600 Park Offices Drive
Suite 150
Research Triangle Park, NC 27709
support@dzone.com
+1 (919) 678-0300

Let's be friends: 🔊 🐦 f in

DZone.com is powered by