This article was first published on TabbForum : Investing in truth: using AI to combat bots and fake news

# Investing in truth: using AI to combat bots and fake news

By Oussama Belmejdoub, Director of Data & Analytics, QuantCube Technology - September 26, 2022

As we approach the US mid-terms, investors are once again trying to predict the results and figure out the potential impact of the elections on the financial markets. We're living in an age in which sentiment on political views, as well as many other topics can be uncovered through the analysis of enormous amounts of data, ranging from traditional sources, such as polls, historical results and expert analysis, to new alternative data sets, such as social media data. But given the proliferation of bots and fakes news, it's essential to have a methodology in place to identify and disqualify such data to avoid the risk of any distortion in predicted results. As a data scientist with extensive experience in building sentiment models using natural language processing (NLP), I'd like to share my approach in uncovering the truth in today's increasingly challenging information landscape.

### The need for validation

When it comes to creating a model for sentiment analysis, the value of alternative data sources cannot be understated. Social media platforms provide a wealth of information that can be analyzed and categorized in real time, whereas traditional means, such as opinion polls and news sources offer snapshots in time which can become quickly outdated.

At the beginning of any sentiment analysis project, it's essential to decide on the data sources that will feed the model, as well as the methodology for creating the sentiment indicators, i.e. ensuring all output is the validated reality. As with any data science project, this requires a long period of discovery, data wrangling and validation – including querying any data variations and making adjustments in line with feedback to improve overall performance. It's also important to ensure all data is anonymized in advance of performing the analysis, in full compliance with data protection and privacy regulations.

Over recent years, QuantCube has proven its expertise in analyzing sentiment data to predict the outcome of upcoming elections. We performed this analysis most recently for the French Presidential elections in April 2022, and previously for the US Presidential elections in November 2020. In both cases our sentiment analysis correlated closely with the final outcome. For example, QuantCube's prediction of Electoral College results published on 1st November 2020 (prior to the elections on 3rd November) was very close to the confirmed results that were published on December 14th. This included accurately predicting  that Joe Biden would win the Rust Belt states (Wisconsin, Michigan and Pennsylvania) that had cost Hillary Clinton's victory in 2016.

### Language matters

For sentiment analysis that focuses on different countries, it's essential that the language of the target population is fully understood. If we are to decide which social media posts are expressing a negative or positive sentiment, we have to ensure that slang and dialect are also taken into account, which can be done at a local, regional, and national level. For these use cases, linguists and native language experts and data scientists are essential.

Ultimately, our job when analyzing social media posts is to identify credible engagements with a topic, such as elections to political office. When it comes to social media platforms, such as Twitter, Telegram and WeChat, a credible source does not have to be an expert, it just has to be a real person engaging with the topic of discussion—but in the age of the bot, this is where things can become difficult.

**Finding fakes**

Increasingly, bots and fake news accounts dedicated to spreading misinformation and disinformation are being used to distort our perception of reality. This is where sentiment indicators that can filter bad information and deliver true insights become invaluable.

I find NLP invaluable for both political indicators and financial indicators – such as our crude oil risk sentiment indicator which analyzes social media sentiment in English and Arabic. For each it's essential to avoid bots and fake news accounts. However, when it comes to political use cases, such as election results, there are far more users—real and fake—engaging with topics, which means we have more data to analyze. In my work creating sentiment analysis indicators, I have found that a significant number of accounts are bots, which must be removed from the data pipeline that feeds our models.

Through NLP, which harnesses the input from subject matter experts and linguists, we can identify these accounts and discount them from the discourse under analysis, i.e. remove them from the indicator. Twitter is, of course, the most popular platform, so I will use this as my example use case.

Deciding which accounts are bots involves a number of stages. Firstly, Twitter provides metadata on accounts, which give us our first layer of analysis. Once suspect accounts are identified through this phase, we process them through our developed similarities algorithms that we trained for several years to confirm those that are bot accounts.

The next layer is where the model must ascribe a sentiment to a Tweet. This requires the creation of a term-document matrix, in which negative, positive, and neutral sentiments can be determined through text analysis. State of the art NLP methods, such as Bidirectional Encoder Representations from Transformers (BERT), can then be used to detect the context, syntax and semantics in text, enabling further accuracy when determining the sentiment related to a subject. This is where working with subject matter experts in advance to analyze key terms and ascribe values is essential.

For an economic indicator, the term "increase production" in a Tweet would be positive when discussing a major exporter of crude oil, but negative when relating to crude oil prices. This is why other terms within the same Tweet must also be considered, as well as the relationship between terms and the context in which they are used. Through an analysis of all the sentiments within a Tweet, our model will provide a score that is either positive or negative—with neutral results discounted from the final output. Now we have a working model that we can continue to fine tune and improve upon.

**No black boxes**

When developing an indicator, it’s essential that the underlying technology, data, and the methodology used to construct the model is entirely explainable. Being able to explain every stage of the process, from data collation and validation to processing and fine tuning, provides confidence to users that the model is not missing key data, was not constructed with bias, and ascribes sentiment in a logical and fair manner.

It's crucial that investors have confidence in the models used. Showing our workings is not only best practice, but is fundamental in delivering continuous improvement and the highest levels of accuracy.

## About QuantCube

QuantCube analyses billions of alternative data points in real time, using artificial intelligence and big data analytics to deliver insights ahead of the market – giving users an edge in their investment strategies. Today we are the global leader in macroeconomic intelligence nowcasting and in pinpointing macro regime change.

Our vision is to become the standard point of reference for macroeconomic, sector, corporate and environmental intelligence. By delivering timely, comprehensive and actionable economic insights we empower users within financial institutions, corporates and public bodies to reach their financial performance and sustainability goals.

Headquartered in Paris, QuantCube employs a diverse international team of economists, quant analysts and data scientists with expertise in multilingual NLP, deep learning and machine learning techniques. The company’s shareholders include Moody’s and Caisse des Dépôts and its R&D in computer vision has been partially funded by the European Space Agency (ESA) and French government space agency CNES.