Big data come from a variety of sources, and among the most popular ones are social media platforms like Twitter and Facebook, to name but a few. Thanks to these networks, businesses now have unparalleled access to huge amounts of data, and this information is used to help detect opinions on specific products or services, analyze consumer behaviours, monitor competitors, and identify early warning signs (EWS). On the basis of these capabilities, it is no wonder that social media analytics has become an important part of competitive intelligence processes in many companies. However, while many of us would agree that data should be an integral part of CI and an essential component of any business strategy, more data isn’t automatically a positive thing. There is, in fact, another side of the story that executives need to be aware of when taking strategic decisions using data gathered from social media platforms.

Several studies suggest that the use of big data from social networks suffer from a series of pitfalls, of which sampling bias and respondent bias are two of the most prominent and overlooked problems. These biases imply that the use of Facebook, Twitter, Instagram and other social media networks data to infer what the general population think might produce skewed results. Why is this so? If you think about it, the first issue is that not everyone uses social media, and of those actively using them, not all are honest when interacting online and offline and may not be entirely represented or even sampled when an analysis is carried out using data from social platforms.

To conclude, while advances in data analytics represent a clear opportunity for businesses, it is crucial to understand that big data from social media might not always provide the whole picture. Business should use a mix of new and traditional data-gathering sources and techniques to minimize bias and extrapolate meaningful and actionable insights. Those relying only on insights extrapolated from social media analytics must be careful to understand how the data was generated, as the risk is to have a huge amount of data…but too little good data!