New application can detect Twitter bots in any language

June 13 2019

Thanks to fruitful collaboration between language scholars and machine learning specialists, a new application developed by researchers at the University of Eastern Finland and Linnaeus University in Sweden can detect Twitter bots independent of the language used.

In recent years, big data from various social media applications have turned the web into a user-generated repository of information in ever-increasing number of areas. Because of the relatively easy access to tweets and their metadata, Twitter has become a popular source of data for investigations of a number of phenomena. These include, for instance, various political campaigns, social and political upheavals, Twitter as a tool for emergency communication, and using social media data to predict stock market prices.

However, research using data from social media data is often skewed by the presence of bots. Bots are non-personal and automated accounts that post content to online social networks. The popularity of Twitter as an instrument in public debate has led to a situation in which it has become an ideal target of spammers and automated scripts. It has been estimated that around 5–10% of all users are bots, and that these accounts generate about 20–25% of all tweets posted.

Researchers of the digital humanities at the University of Eastern Finland and Linnaeus University in Sweden have developed a new application that relies on machine learning to detect Twitter bots. The application is able to detect autogenerated tweets independent of the language used. The researchers captured for analysis a total of 15,000 tweets in Finnish, Swedish and English. Finnish and Swedish were mainly used for training, whereas tweets in English were used to evaluate

the language independence of the application. The application is light, making it possible to classify vast amounts of data quickly and relatively efficiently.

"This enhances the quality of data—and paints a more accurate picture of the reality," Professor of English Mikko Laitinen from the University of Eastern Finland notes.

According to Professor Laitinen, bots are relatively harmless, whereas trolls do harm as they spread fake news and come up with made-up stories. This is why there's a need for increasingly advanced tools for social media monitoring.

"This is a complex issue and requires interdisciplinary approaches. For instance, we linguists are working together with machine learning specialists. This type of work also calls for determination and investments in research infrastructures that serve as a platform for researchers from different fields to collaborate on."

According to Professor Laitinen, it is essential for researchers to have access to social media data.

"Currently, data are the property of American technology conglomerates, and a source of their income. In order for researchers to gain access to this data, cooperation at the national and international levels, and especially the involvement of the EU are needed."

 More information: Towards a language independent Twitter bot detector www.diva-portal.org/smash/record.jsf?pid=diva2%3A1302270&dswid=-9446