Sentiment Analysis-The Emotionality of Discourse

Mike Thelwall, University of Wolverhampton, UK

The goal of this workshop is to apply the sentiment analysis software SentiStrength to classify the sentiment of some YouTube comments and then read the very strongly positive or negative comments to discover (a) what strong sentiment is expressed about and (b) how the magnitude and topic of sentiment differs between groups or videos. The session will start by describing how SentiStrength works (15 mins) and then what the prepared Python script does (5 min). Participants can then try out the Python scripts on their datasets.

Workshop Instructions

The first part is to get everything working on your computer and the second part is exploration.

Getting SentiStrength working on your computer

All the software and data files are in the following folder http://sentistrength.wlv.ac.uk/jkpop/.

1. Make a new folder on your computer to save the software and data.
2. With a web browser, download the SentiStrength program **SentiStrength.jar** to the new folder on your computer.
3. Download the SentiStrength linguistic data files **SentStrength_Data.zip** to the new folder on your computer and unzip it. Make sure that the linguistic files are inside a separate folder.
4. Download the YouTube data files **TWICE-BTS-EXO-BLACKPINK_english comments, one file per video.zip** and extract one of them from the zipfile.
5. Download the Python program **ClassifyCommentSentiment.py** to the new folder on your computer. Open this program and edit the lines below to point to the location of SentiStrength, its data files, and the YouTube comment file on your computer.
    a. SentiStrengthLocation = "D:/Downloads/SentiStrength.jar" #This must point to the location of SentiStrength on your computer
    b. SentiStrengthUnzippedTextFilesLocation = "D:/SentiStrength_Data/" #This must point to the location of the unzipped SentiStrength data files on your computer
    c. FileToClassify = "E:/data/YouTube/BTS/BLACKPINK_eng-_NVwS4mcVYg_commentsOnly.txt"
6. Run the Python program and look at the results – sentiment classifications for the file. If it does not work, check the file and folder paths above.

Exploration tasks

Each file in in the large zipfile you downloaded contains all the comments on a single video. The start of the filename gives the name of the group. The funny string of characters is the YouTube ID of the video. Add this string to the end of https://www.youtube.com/watch?v= to get the video URL. For example, the file **BLACKPINK_eng-FzVR_fymZw4_commentsOnly.txt** is from BLACKPINK video https://www.youtube.com/watch?v=FzVR_fymZw4.

Extract one large file of video comments from each of the different groups in **TWICE-BTS-EXO-BLACKPINK_english comments, one file per video.zip** and use the instructions above to classify them for sentiment. The numbers at the start of each line are the strength of positive (on a 1 to 5 scale) and negative (on a -1 to -5 scale) sentiment.

Load the four _out.txt files with sentiment results into a spreadsheet and then read the very strongly positive or negative comments to discover (a) what strong sentiment is expressed about and (b) how the magnitude and topic of sentiment differs between groups. Are your findings common sense or unexpected?

If you have time, use the spreadsheet to calculate the average positive and negative sentiment strength for each video and compare the averages between positive and negative sentiment and between videos.

About SentiStrength

SentiStrength is a commercial sentiment analysis Java program that available free for academic research. It reads texts and then assigns them two scores: one each for positive and negative sentiment strength.

-1 (no negative sentiment) to -5 (contains extremely negative sentiment)

1 (no positive sentiment) to 5 (contains extremely positive sentiment

SentiStrength has a dictionary of over 3,000 words that are pre-classified for sentiment strength (e.g., love=+3; hate=-4) and applies these scores to words when found in a text. It uses rules to cope with sentiment expressed or modified in other ways, such as negation (e.g., not happy), boosting (very nice!), emoticons :) and sentiment spelling. Yaaaaaay!!!

SentiStrength can be tried out online at: http://sentistrength.wlv.ac.uk and is described and evaluated in the following journal articles, both of which are available free online.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), 163-173.

For Java users, there is a bit of extra information for the Java version on the SentiStrength website. There is no JavaDoc but there is a brief manual online. http://sentistrength.wlv.ac.uk/documentation/SentiStrengthJavaManual.doc

About Mike

Mike Thelwall is a professor of information science and leads the Statistical Cybermetrics Research Group at the University of Wolverhampton, UK. He has developed and evaluated free software and methods for systematically gathering and analysing web and social web data, including for sentiment analysis. He is the developer of the SentiStrength software. He has co-authored hundreds of refereed journal articles and five books, including “Social web research with Mozdeh” (free online). CV: http://www.scit.wlv.ac.uk/~cm1993/mycv.html
Mugshot: Picture: http://www.altmetrics.ntuchess.com/image/MikeBoat.png