USE OF SENTIMENT ANALYSIS TO ASSIST POLICE INVESTIGATIONS

A Policy and Practice Briefing from the Digital Forensics and Social Media project funded by the Dawes Trust

Authored by:
Dr Martyn Harris
Dr Janice Goldstraw-White
Iacopo Pozzana

September 2022

UCL

perpetuity research

Birkbeck
UNIVERSITY OF LONDON

ICPR
Institute for Crime & Justice Policy Research

Copyright

Acknowledgements

This paper forms part of a larger project on Digital Forensics and Social Media (DFSM) made possible by the generous support of the Dawes Trust, to whom we are extremely grateful.

The authors would particularly like to thank my colleagues on the project team, Professor Martin Gill, Professor Jessica Jacobson, Professor Ben Bradford, Tiggey May, Dr Kari Davies, and Owen Wilkie for their support and feedback. In addition, they would like to extend their gratitude to those who made comments on earlier versions of this paper, whose feedback and insights were invaluable. In particular, we would like to thank Professor Alessandro Provetti.

The views expressed in this paper are solely those of the authors.

What is sentiment analysis?

Sentiment analysis is an analytical technique involving the analysis of people's opinions, sentiments, evaluations, attitudes, and emotions from written or verbal language. It is also known as opinion mining, text/web mining, subjectivity analysis, appraisal extraction or emotion artificial intelligence (AI). More specifically, sentiment analysis describes a family of natural language (NLP) and machine learning approaches, which classify words and phrases in a sentence, paragraph, or document, according to whether they express a positive, negative, or neutral opinion, known as polarity.

The use of such analysis has risen with advances in computer and information sciences alongside the growth of social media, and particularly the commercial use of feedback and review sites, which provide huge volumes of opinion data in digital forms. Opinions are central to nearly all human activity and therefore are a key influence on behaviours.[i] Sentiment analysis is employed in the marketing sector to track opinions and sentiment towards products and brands, for example through application to Amazon reviews. Researchers also apply this approach to analysing political opinion during general elections.

Sentiment analysis is not new, however. Over the last two decades it has developed from a relatively simple analysis system that looked for and counted the frequency of particular words, to systems that now not only look at the words being used, but also consider the context in which they appear, giving a more nuanced picture of the sentiment behind them. The increased use of AI and machine learning has greatly aided the development of sentiment analysis.

Why we think that sentiment analysis could be useful to the police

Year on year, an ever-greater proportion of social interactions are taking place online, producing vast amounts of unstructured textual data. This offers the potential for valuable insights to be revealed through the application of statistical analysis to these data.

In police investigations, digital communications between suspects, witnesses and complainants can generate important information and evidence. However, handling such data poses significant challenges due to its sheer volume, as well as the complexity arising from the use of paraphrasing, spelling errors, abbreviations, and slang terms. Current police approaches to the analysis of such data appear to be limited to manual reviews, which are slow and painstaking, often coupled with broad keyword searches.

In contrast, sentiment analysis has the potential to offer swifter and less resource-intensive techniques whereby text-based digital communications can be transformed into a time series in which “events of interest” can be identified, and portions of the exchanges highlighted for more detailed attention.

How has sentiment analysis been used in research and crime studies before?

Sentiment analysis has not been as widely used in social policy and the social sciences as it has in the customer and retail sectors. However, there have been some good examples of application of this approach to crime issues, for example with respect to:

- Examining public opinion on crime[ii]
- Detecting crime behaviour patterns[iii]
- Analysing online extremist behaviour[iv]
- Using mobile phone networks to investigate organised crime structures[v]
- Learning to predict intimate partner violence[vi]

There have also been a number of studies which deal with the use of sentiment analysis for purposes of 'content detection' on social media, in conjunction with author identification. This involves identifying the type of language encoded by a text, for example, whether it contains abusive or threatening language. In addition, texts can also be grouped by author through analysis of vocabulary choices and the distribution of words and punctuation. Some example applications of these techniques include:

- Harmful content detection[vii]
- Indicators of so-called "darker traits of human personality"[viii]
- Measuring hate speech and white nationalist rhetoric[ix]
- Presence of youth gangs on social media[x]

What kinds of data can be subject to sentiment analysis?

Sentiment analysis can be applied to any natural language text data, including that sourced from:

[i] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

[ii] Prichard, J., Watters, P., Krone, T., Spiranovic, C., & Cockburn, H. (2015). Social media sentiment analysis: A new empirical tool for assessing public opinion on crime? *Current Issues in Criminal Justice*, *27*(2), 217-236.

[iii] Bolla, R. A. (2014). *Crime pattern detection using online social media*. Missouri University of Science and Technology.

[iv] van der Vegt, I., Mozes, M., Gill, P., & Kleinberg, B. (2019). Online influence, offline violence: Linguistic responses to the'Unite the Right'rally.

[v] Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications*, *41*(13), 5733-5750.

[vi] Petering, R., Um, M. Y., Fard, N. A., Tavabi, N., Kumari, R., & Gilani, S. N. (2018). Artificial intelligence to predict intimate partner violence perpetration. Artificial intelligence and social work, 195.

[vii] Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., & Ungar, L. (2016, October). Studying the Dark Triad of personality through Twitter behavior. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 761-770) and Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012, December). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th international conference on machine learning and applications* (Vol. 2, pp. 386-393). IEEE.

[viii] Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of research in personality*, *36*(6), 556-563.

[ix] Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., ... & Tucker, J. A. (2018). Measuring the prevalence of online hate speech, with an application to the 2016 US election.

[x] Blevins, T., Kwiatkowski, R., Macbeth, J., McKeown, K., Patton, D., & Rambow, O. (2016, December). Automatically processing tweets from gang-involved youth: towards detecting loss and aggression. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2196-2206).

- Digital archives
- Online communities
- Discussion boards
- Weblogs
- Product rating sites
- Messaging apps
- Chatrooms
- Price comparison portals
- Newsgroups

How does sentiment analysis work?

Performing sentiment analysis encompasses several tasks. In the context of machine learning, a statistical model is constructed and trained on a collection of data. This model is then used to predict a label, score, or other type of output, such as a label defining the genre of the text, a sentiment score, or a translation of the text into another language. The aim of machine learning is to produce a model that is accurate, but also generalisable so that it will continue to produce accurate results when encountering data, it has not been trained to model. Natural language data is often unstructured and complex, with users adopting a range of lexical devices (syntax), and vocabulary (synonyms) to define or discuss the same things. The challenge with machine learning applied to natural language data, is to group instances of the same topic together, either for classifying documents or identifying only negative texts.

Before training a machine learning algorithm, we must provide the model with the correct answers on which it can extrapolate when presented with new and previously unseen data. In the context of sentiment analysis, human annotators are often tasked with assigning a score to each sentence of a dataset according to a set range, for example, a five-star rating commonly used in product reviews. Typically, annotators would be tasked with annotating on average 250–500 sentences for each sentiment polarity (negative, neutral, positive). In addition, multiple annotators can also be used if the data is large in scale, and their level of agreement can be assessed to help reduce bias and to obtain high quality annotations for training the sentiment classifier model.

Training a machine learning algorithm to identify the overall sentiment of data reflected by natural language texts involves partitioning the complete collection of annotated data into a training set (80% of the data) and a test set (20% of the data), each composed of sentences and their associated sentiment score given by human annotators. The training set is used to 'teach' the model the correct classification of the data. During this process, the algorithm will compute and adjust the score of the words in each sentence in the training dataset according to its polarity given by the annotators. More specifically, a binary sentiment classification involves labelling opinions as expressing either an overall positive or negative opinion, whereas a multi-class sentiment classification involves locating the sentiment by using one of five categories: strong positive, positive, neutral, negative, or strong negative.[xi] A further approach assigns a score to the sentence or phrase. In this instance, the sentiment model assigns a score of 0 when classified as neutral, negative if it falls below 0, and positive if the score is above 0. The overall range of sentiment is between -1 and +1. On completion of the training phase, the model is then used to make predictions based on the sentences in the test set, on which it was not trained, which allows us to compute how accurate the model is in achieving its task.

The resulting model provides the mechanism for analysing the dynamics of sentiment in a conversation between individuals, or in the response of social media to a particular brand or product. In the context of a police investigation, analysing sudden shifts in the sentiment score may provide the means to reduce a large body of digital information to a subset of relevant information to aid a criminal investigation, for instance, focusing on a subset of texts that are considered very negative in sentiment or opinion.

A further application of machine learning techniques, which is complementary to sentiment identification, is to establish the target entity of the sentiment, that is who the sentiment or opinion is directed at. This form of analysis is known as Named Entity Recognition (NER), nouns describing people, places and things. In a criminal investigation, this would help to identify texts composed of words related to commodities including drugs and weapons; people, including the name of the victim and the accused; and locations pertaining to the crime scene or areas under investigation.

A pilot study to inform use of sentiment analysis in police investigations

We undertook a pilot study involving the application of sentiment analysis to two sets of communications. Each set of communications was over a period of more than a year and between two young people who knew each other well; one was via WhatsApp and the other SMS. The aim of this work was to devise a preliminary framework for use in development of investigative tools that could be used by the police to facilitate analysis of large volumes of messages, sent over periods of months or years between individuals. Such tools would provide an alternative to manual and keyword-based reviews of this kind of material.

For the pilot, we developed a framework based on the application of sentiment and time series methods[xii] to the analysis of the two sets of digital communications. The framework supports the analysis and visualisation of message volumes and sentiment over time to allow "events of interest" to be identified. These events are defined as periods of the volume or sentiment time series that are novel or unexpected in reference to what has preceded them. For example, there may be a period of increased negative sentiment in messages sent between individuals, followed by a sudden decrease in message volume, which would suggest a breakdown in communication. In the context of a police investigation, these anomalous events would help to reduce the message data to a smaller subset that could be more quickly and effectively reviewed when time is critical.

The framework comprised a set of approaches, including sentiment analysis, named entity recognition, message volume and measures of lexical richness (the quality of vocabulary used by participants). In combination, these support the development of software tools to aid criminal investigators with the processing and analysis of digital information from mobile devices and social media platforms at scale. In our pilot study we explored whether the analysis of message volume, sentiment, lexical diversity, and named entities provided any indication of the dynamics of the relationship (such as a temporary breakdown in that relationship), the impact of events (such as difficulties in a relationship between a participant and a third party, or a traumatic external experience), and topics of interest to the participants (such as mutual associates, locations and commodities).

From our preliminary analysis of the data, we were able to identify for each set of communications and over various time periods:

- The daily and weekly average volume of messages sent by each participant in each group, in order to create a baseline for identifying sudden or short-term increases or decreases in message frequency.

- The daily and weekly trend in sentiment for each participant in each group. This analysis classified messages on the scale of -1 and +1 to classify sentences as either positive, negative or neutral, where a sentiment was expressed, and identified daily averages and trends to identify periods of change in sentiment over the full period of interaction.
- The lexical diversity of the text. Linguistic evidence is already used in criminal investigations and proceedings and analysis of lexical diversity would contribute to this, by identifying messages and portions of exchanges which are verbose and information rich compared to those that are terse. This, in turn, could assist author detection or identification of changes in the tenor of communications.
- Named entity recognition (such as individuals' names, locations, organisations, time expressions, values and so on). Such analysis can be further filtered and used to direct investigators to subsets of messages which pertain to certain individuals, locations, or commodities such as weapons or drugs.

We presented a case study based on mobile messaging data provided by two groups of participants who had regular communication and knew each other on a personal level. The main contribution is a framework for developing semi-automated digital tools for summarising unstructured language data through time series analysis applied to the output of NLP methods such as sentiment and NER. The aim is to guide investigators towards information stored in digital format to facilitate the gathering of evidence for criminal investigations.

The results of the analysis of messages provided by our first group of participants revealed two events in the daily trend in volume, marked by a sudden increase in messaging leading up to both events. We also observed a long-term decline in both sentiment and lexical richness as the communication between participants continued by looking at a weekly trend, which suggests that a final event late in the conversation may have had an impact on subsequent communication volumes and sentiment. This may potentially indicate a breakdown in the relationship between participants reflected by a drop across our measures. Similar results were observed for the second group of participants, but due to less interaction, it was not possible to generate a long trend over the measures. Consequently, when interaction is low it is not always possible to apply all the methods introduced, though short-term trends may still provide some insight.

The results of applying NER provides a means to detect novel terms such as drug names, track movements of individuals through mention of location entities in their messages, as well as interactions with third parties, through the extraction of named entities related to people and organisations. Furthermore, the analysis of volume, sentiment, NER, and lexical diversity measures could be useful in isolation, or in combination, to identify dominant relationships, a breakdown in a relationship, or cyber-bullying.

The approaches we outline in our study require relatively little data compared to machine learning approaches based on Neural Networks, which are considered state-of-the-art, but in fact require more human effort in tuning the parameters of the model in order to produce accurate results. However, we acknowledge that the accuracy of machine learning sentiment models and long-term trends are not always possible with very small or sparse datasets represented by infrequent interaction.

The framework developed in our study is composed of a set of approaches that could be combined to develop software tools and dashboard interfaces to aid investigators in a criminal investigation with processing and analysing digital information from mobile devices and social media platforms to reduce the manual work needed to identify regions of interest, defined by a subset of digital texts from a potentially large data set that are relevant to an investigation. This is particularly so when the data obtained can span several years with regular exchanges between multiple participants, making it difficult for an investigator to investigate when time is critical.

xi Bolla, R. A. (2014). *Crime pattern detection using online social media*. Missouri University of Science and Technology.

xii Time series methodology records data at regular intervals over a set period of time