Forensic Science International: Digital Investigation 49 (2024) 301753

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Sentiment and time-series analysis of direct-message conversations

Martyn Harris, Jessica Jacobson, Alessandro Provetti *

Birkbeck - University of London, Malet Street, London, WC1E 7HX, UK

ARTICLE INFO

Keywords:
Text analysis
Sentiment analysis
Access to mobile data
Digital forensics

ABSTRACT

Social media and mobile communications in general are an extremely rich source of digital forensic information. We present our new framework for analysing this resource with an innovative combination of time series and text mining methods. The framework is intended to create a tool to analyse and operationally summarise extended trails of social media messages, thus enabling investigators for the first time to drill down into specific moments at which sentiment analysis has detected a change of tone indicative of a particularly strong and significant response. Crucially, the method will give investigators an opportunity to reduce the time and resource commitment required for ongoing and hands-on analysis of digital communications on media such as Texts/SMS, WhatsApp and Messenger.

1. Introduction

Digital forensics involves the extraction of information retrieved from digital systems, which is then processed and explored in order to elicit intelligence for the purpose of police investigations, or as evidence in criminal proceedings, Tully et al. (2020). Due to rapid technological advancement more of our social interactions take place in an online setting through digital means. As a consequence, criminal investigations are seeing a reduction in the quality of digital forensic results, Casey (2019) and Krishnan et al. (2022), due to the complexity of the data, whilst methods and approaches for analysing digital evidence obtained from websites, storage media, laptops and mobile devices, are still being developed.

Today a sort of second-level Digital forensics is coming into demand: one that assumes that data has been obtained/extracted (which is the first level, where the most research is) and focuses on extracting insight from the data. The case at hand is that of 'conversations' over SMSs, WhatsApp, Messenger, WeChat and others. When investigations focus on people who are closely related, their SMS exchanges might go back several years and involve hundreds of messages per day, use of jargon, emojis and other specific text analysis challenges. Even in a best-case scenario where the mobile device is available, the data is extracted cleanly etc. investigators are faced with the sheer volume and challenges of such conversations.

Semi-automated 'text mining' approaches can aid investigators by highlighting patterns, sudden or unexpected changes in dynamics expressed by outliers as a means to filter a potentially large data set spanning several years of communication to the most relevant for the investigation. This could facilitate a Police investigation by providing investigators with an opportunity to reduce the time and resource commitment involved in manually sifting through SMS conversations which are likely to be long-term and with an high frequency of exchanges. Beyond the investigation of family/friendship links, Holt et al. (2015) describe several investigative domains where text-oriented digital forensic tools can also assist, including cyber-crime (see Coyac-Torres et al., 2023 for an approach based on neural networks), bulling, stalking, terrorism and extremism.

In this article We present our pilot framework based on the application of time series methods to the analysis of textual communications between groups of individuals using mobile phone message platforms, including *WhatsApp and* SMS. Our interpretation model is based on time series, so to accommodate the time element of investigations, where we aim to construct a time-line of events leading up to, onset, and conclusion of a crime.

The ultimate aim of our research is to prove that semi-automated methods can assist in identifying events of interest, represented by a subset of messages sent by individuals, that reflect a change in the dynamics of the relationship or the impact of external events on the sentiment encoded by the messages. However, a direct application of the recent advances of Natural Language Processing (NLP), notably Large Language Models, is beyond the scope of this work and, arguably, not yet advisable as it would raise technical, ethical and regulatory issues that are still under discussion at a general level and unlikely to be solved soon. So, differently from recent NLP trends, e.g., Studiawan

* Corresponding author.
E-mail address: a.provetti@bbk.ac.uk (A. Provetti).

https://doi.org/10.1016/j.fsidi.2024.301753
Received 18 June 2023; Received in revised form 18 April 2024; Accepted 20 April 2024
Available online 20 May 2024

et al. (2020) which addressed 'forensic timelines' over logs, our framework does not rely on LLMs and in general minimises the training effort for the Machine learning core component, which will be described below.

Semi-automated text mining approaches provide investigators with an opportunity to reduce the time and resource commitment involved in current the manual analysis of texts, particular when time is limited.

To process long-term SMS conversations we have developed several methods for exploring trends in social relationships between individuals who know each other well on a personal level. We explore whether the analysis of message volume, sentiment, lexical diversity, and named entities provides any indication of the dynamics of the relationship, including the impact of external or third-party events e.g. the breakdown of a relationship between participants or a third party, as well as topics of interest around people mentioned outside of the relationship, locations, and commodities.

In the remainder of the paper, we introduce the mobile data used for our study in Section 2. Next we introduce the methods, in Section 3, which were developed for pre-processing and summarising the textual data. We describe well-established time series methods such as first-order-differencing and moving average to identify trends. We also discuss approaches in Natural Language Processing (NLP) including sentiment analysis, Named-Entity Recognition (NER) and measures of lexical diversity (LD), which are summarised using time series methods. In Section 4, we present the results of the time series analysis and discuss our findings. We conclude the presentation of our results with Section 5, and discuss further work that will be explored in Section 5.1. In the next section, we describe the data used as the basis for the analysis.

2. The data

The data used for our analysis is derived from two data sets which were collected in 2022 in the framework of our externally-funded project entitled "Digital forensics and social media: Challenges and opportunities for law enforcement," which was awarded by the Dawes Trust, a UK charity that sponsors research in the forensics sciences. The project focused on engaging with UK police corps and the Crown Prosecution Service for England and Wales (CPS) to elict their current needs in terms of software to supports digital forensics investigations.

The key aspect of our data collection is the involvement of four volunteers who donated, under a strict confidentiality/non-disclosure agreement, the data needed for this study. They are two pairs of close friends, now in their late teens/early twenties, who maintain a lively, continuous contact. One pair use SMSs while the other uses WA. Participants were freely exchanging messages for a long period and only afterwards were contacted to check their availability to participate in the study. This is excludes observation bias. Their closeness and frequency of contact make their message trail a good example of what data an investigation over British young adults might work on.

The confidentiality of the donated data unfortunately hinders the full reproducibility of the analysis. However, we believe that, unlikely as we are to find someone willing to share their most intimate conversations with the research community, it is in fact coherent with the ultimate goal of our project: design and test a solution for real, unfiltered, non-anonymous conversation texts that are similar to that collected and handled by investigators.

The first dataset consists of a collection of 46,304 messages exchanged between two participants, p_1 and p_2, using *WhatsApp*, over a period of 482 days. The second dataset is composed of 38,920 *SMS* messages sent between participants p_3 and p_4 over a period of 405 days.

Before proceeding with the analysis, we anonymise references to the participants in the data and refer to those classed as *group 1* as p_1 and p_2, and for *group 2* we define them as p_3 and p_4. We also anonymise personal names mentioned in any messages presented in the discussion.

Table 1
Summary of the data collected from each group.

	Group 1	Group 2
Platform	WhatsApp	SMS
Duration (days)	482	405
Total (messages)	46,304	38,920

Table 2
Example sentences from our WhatsApp and SMS datasets.

Examples
It's just I rly rly don't want to do something that's gonna make me uncomfortable and make things awkward for her tooo
Ohh shit idk
Bc that was even before I suggested not going
Because yano... he's actually being paid for his haha
wdym (What Do You Mean), also *wym* (what you mean?)
tysm (thank you so much)

In Table 1, we present a summary of the data collected for analysis together with basic statistics.

2.1. The specificity of 'SMS-speak'

Social media data can be noisy and complicated to process. This is due to the presence of punctuation and emojis for emphasis, but also due to the conventions in spelling adopted by short text message style social media posts. Historically, mobile phone text messages have been restricted to no more than 140 characters. Other social media platforms, such as Twitter, impose similar restrictions on message length. As a result, these limitations have influenced how we communicate and pack in the semantic information being conveyed in to limited characters. Some mechanisms include the use of abbreviations and acronyms, which when combined, make social media text data challenging to preprocess and analyse with accuracy, Hussein (2018). Some examples of abbreviations and acronyms are presented in Table 2. Personal names mentioned in the text have been reduced to the first character, and highlighted in italics.

3. Methods

Our objective is to identify a subset of messages exchanged between individuals over a period of time that might point to conflict, disagreement or sudden changes in the mood or nature of the relationship. The determination of the specific time intervals were such changes are detected are the essential output of our analysis as they will help investigator in focusing their analysis on specific moments and events.

Time-series methods are well-suited to the task described above since an investigator can explore a time series to find messages related to an investigation, either before or immediately after a crime has been committed. We define the subset of messages to be explored as an *event of interest,* which may be determined by above average message volumes, extreme changes in 'polarity' (defined later) through sentiment analysis, and changes in lexical diversity, which may indicate a shift in how participants are communicating. Named-entities recognition is also a useful component, since the investigator can quickly determine whether locations or people involved in a crime, are also mentioned in messages during the period leading up to and after a crime has been committed.

We apply a time series approaches, defined, e.g., in Chatfield (2004), to the analysis of WhatsApp and SMS messages based on the trend in volume, sentiment, and lexical diversity. We also apply Named Entity

Recognition (NER), Nadeau and Sekine (2007), to explore names of people and locations in the generated time series.

The sentiment analysis and named-entity recognition was performed using our NLP suite Samtla,[1] Harris and Levene (2021), which is a framework for annotating digital texts with sentiment. In Samtla, named entities are discovered using semi-supervised techniques, which in fact required us to manually annotate a small subset of the messages. Samtla also does sentiment annotation by the *pSenti*[2] pre-trained model of Mudinas et al. (2012, 2018).

Samtla and its pSenti analyser were developed to support NLP options, and sentiment analysis in particular, in linguistic domains which have little in terms of annotated corpora or even sources. I.e., where traditional ML techniques are unlikely to work due to lack of training. Whereas the standard approach in the literature is to *port* models trained from a data-rich domain, pSenti works in-domain with a semi-supervised method. Starting from manual annotations a few typical sentiment words (*seeds),* pSenti performs vectorisation and linear SVM classification to create a domain-specific sentiment lexicon. Experiments in, e.g., Mudinas et al. (2018) have shown that this solution works better when applied in a stratified method: boosting and SVM at lexical level followed by boosting and LSTM at document level. The relative complexity of pSenti training pays off in terms of direct applicability and classification ability even against fully-supervised solutions, Harris et al. (2024).

We continue this section with an introduction to the time series methods adopted for analysing trends in volume, and the sentiment, lexical diversity, and named entities encoded in the message texts.

3.1. Time series analysis

We analyse the volume and sentiment of messages sent over time between each participant and group in order to identify periods of frequent and infrequent exchange, as well as sudden changes in volume and sentiment of the messages, which might suggest an event occurred resulting in an increase in the rate of communication over a short period.

To achieve this, we apply a number of well-motivated time series methods, Chatfield (2004), including a moving average to identify trends over short and long periods of time, and Exploiting such a clustering structure, we are able to utilize machine learning algorithms to induce a quality domain-specific sentiment lexicon from just a few typical sentiment words (“seeds”). An important finding is that simple linear model based supervised learning algorithms (such as linear SVM) can actually work better than more sophisticated semi-supervised/transductive learning algorithms which represent the state-of-the-art technique for sentiment lexicon induction. The induced lexicon could be applied directly in a lexicon-based method for sentiment classification, but a higher performance could be achieved through a two-phase bootstrapping method which uses the induced lexicon to assign positive/negative sentiment scores to unlabelled documents first, a nd t hen u ses those over the moving average time series to identify the periods of exchange with high and low activity.

The moving average involves applying a sliding window over the time series. The moving average measures the stability of the time series, and provides a form of smoothing to reveal the trend in the data over the period in question. The moving average has a window parameter n, which determines the number of consecutive observations per window, in our case a fixed window of 24 hours and 168 hours to obtain a daily and weekly trend, respectively. For each window of n values, we compute the mean, resulting in a new smoothed time series for the given period. We note that the larger the window size the more smoothing is applied to the time series to produce trends over different periods of length e.g. weekly, monthly, annually.

To mitigate against the potential lose of data when setting large window sizes, we calculate a centred window, where the current value is the middle value of the window, padded with an equal number of historic, and future observations on either side.

First-order differencing involves calculating the difference between the current value of the time series and the value of the previous time point. The resulting plot will reveal where there were large changes in the volume or sentiments.

We apply these approaches to the volume and sentiment of the messages after generating the time series. This involves computing the total number of messages sent, and the average sentiment on an hourly basis. We then apply a moving average with a fixed window of 24 hours (daily trend) and 168 hours (weekly trend), respectively to the volume and sentiment time series for each participant. We also compute the first order difference over the resulting moving average for each time series.

A last step, involves calculating the average volume of messages sent by each participant for the whole conversation period to act as a baseline to identify when participants increased or decreased the rate of messaging.

Table 3
A sample of the words listed in the positive lexicon used for training the sentiment classifier.

‘yesss’
‘Yeahh’
‘gooodd’
‘lol’, ‘Loool’, ‘loool’, ‘looool’
‘haha’, ‘hahaha’, ‘hahahah’, ‘Hahahah’, ‘hahahaha’
‘tysm’
‘wowww’

3.2. Sentiment analysis

A further component of the study applies Sentiment Analysis techniques (see, e.g., Liu (2015)) to each group of messages. Mobile text messages are often short and composed of abbreviations that are problematic when applying sentiment analysis techniques since many words will be out of vocabulary resulting in unreliable sentiment scores. We adopt the same semi-supervised approach of Mudinas et al. (2012, 2018), i.e., we augment a sentiment classifier trained on Twitter data, with further training data consisting of sentences extracted and manually compiled from our data set to tailor the model to the domain of the texts. In addition, by using a base model trained on a larger data set, we can mitigate against a lack of mobile message data on which to reliably train a sentiment classifier. We also present a small lexicon of domain-specific words to assist in tailoring the domain of the resulting sentiment classifier, which are presented in Table 3 and 4.

We evaluate trained sentiment classifier on a manually-annotated test set composed of 300 text messages, which were randomly sampled, without replacement from the messages provided by both groups of participants. To evaluate the model, we composed a test set of 100 positive, neutral, and negative examples, and compute the precision, recall and F_1-score of .74 achieved by the model on our test set. (See Table 5.).

The F_1-scores revealed that the model performed best at identifying messages with negative sentiment (.76), followed by positive (.75) and, finally, neutral (.69). We score all messages using the trained model, and compute the average, sentiment on an hourly basis. We apply a moving average with a fixed window of 24 hours representing the trend in sentiment over the period of a day for the volume of messages. We also compute the average sentiment for the whole conversation for each participant which we treat as a baseline (see Fig. 1).

[1] Samtla stands for *Search And Mining Tools for Language Archives* and it was accessed via the *SamtlaAPI*, available at samtla.dcs.bbk.ac.uk/samtlaAPI/ upon registration.

[2] The pSenti library is also available from github.com/AndMu/Wikiled.Sentiment.

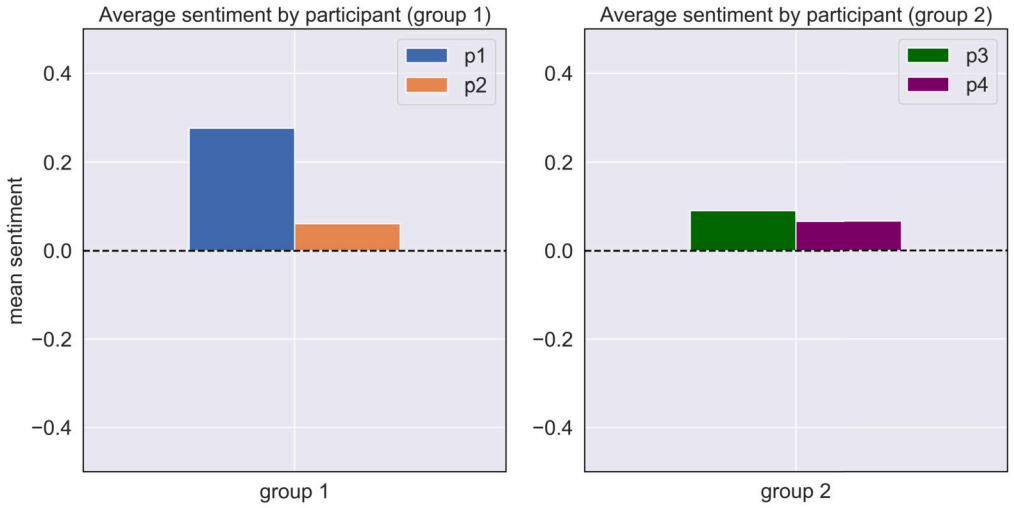

Fig. 1. Plots of the average sentiment for the whole conversation by group and participant, with group 1 presented on the left and group 2 presented on the right.

Table 4
A sample of the words listed in the negative lexicon used for training the sentiment classifier.

'noo', 'nooo', 'noooo'
'wtf'
'af'
'shittt'
'ffs', 'Fffffffs'
'badd'

Table 5
The overall F_1-score performance scores for the sentiment classifier, based on the test set of 300 positive, neutral, and negative examples.

	Precision	Recall	F_1
Score	0.841	0.680	0.752

We consider this a measure of an individuals' disposition, in other words, whether they are generally optimistic or pessimistic in the sentiment of their messages. This will enable us to identify when the sentiment of individuals fluctuates away from what we regard as their individual baseline.

3.3. Lexical diversity (LD)

As the name suggests Lexical Diversity (LD), Torruella and Capsada (2013) is a measure of how many unique words there are in a text. Lexical words are defined as words falling into the category of nouns, adjectives, verbs, and adverbs, which convey the meaning of a text. Lexical diversity provides information about a language user, including their proficiency with the language (native versus second language learner) and can also provide clues as to their age (language acquisition).

There are several measures available for analysing LD, Fergadiotis et al. (2013), however, some measures can be sensitive to text length, which can produce inaccurate results. We apply the Moving Average Type–Token Ratio (MATTR) measure of LD, Covington and McFall (2010), which is a non-parametric measure for assessing the breadth of a speaker's vocabulary from a language sample, and is empirically well-motivated for producing unbiased measures of lexical diversity compared to simpler approaches based on the Type–Token Ratio (TTR), Fergadiotis et al. (2013). Furthermore, MATTR is considered appropriate when the aim is to identify dysfluent production, Covington and McFall (2010), or in our case short versus verbose messages sent by participants in each group. The measure is obtained by calculating the TTR over a moving window of a fixed length, where a window size of five or greater ensures the MATTR is uniform. We also compute the MATTR for each participant with a fixed window of 1000 words to obtain a baseline of lexical diversity for each participant.

To visualise the results, we adopt several time series methods to analyse the volume, sentiment, and lexical diversity of messages for each group of participants. More specifically, we apply a moving average, with a window fixed at 24 and 168 hours, respectively, representing the daily and weekly trends. We also generate a time series based on first-order differencing, Chatfield (2004) to reveal daily and weekly differences over our chosen metrics of volume, sentiment, and lexical diversity.

3.4. Named entity recognition

We apply Named Entity Recognition (NER) in order to extract named entities representing people, locations, organisations and commodities from the messages.

We adopt an approach based on *Conditional Random Fields* (CRF), described, e.g., in Sutton and McCallum (2012) as a probabilistic graphical model for sequential data, and predict the named entity label (person, place, commodity, organisation etc.) for words in a sentence. A CRF predicts the label for a given token by taking advantage of its surrounding context encoded as features, Lafferty et al. (2001). We design a set of feature functions to extract features for each word in a sentence from the training data.

We train the CRF with the well-known *Gradient Descent* method. In particular, the so-called Limited-memory Broyden-Fletcher-Goldfarb-Shannon method (LM-BFGS) was the starting point of our training. During its training, the CRF model estimates and adjusts the weights of each feature function so that to maximise the likelihood of the labels given the training data. The process for training is as follows:

1. We tag words relating to named entities from a small subset of the message data (200 sentences) to act as the training data for the baseline model.
2. We train the model and predict the tags for the current training data.
3. We then compute the probability of the label sequence for the current input sequence, and select the top-m Viterbi sentences with a probability equal to or higher than a predefined threshold (> 0.9).

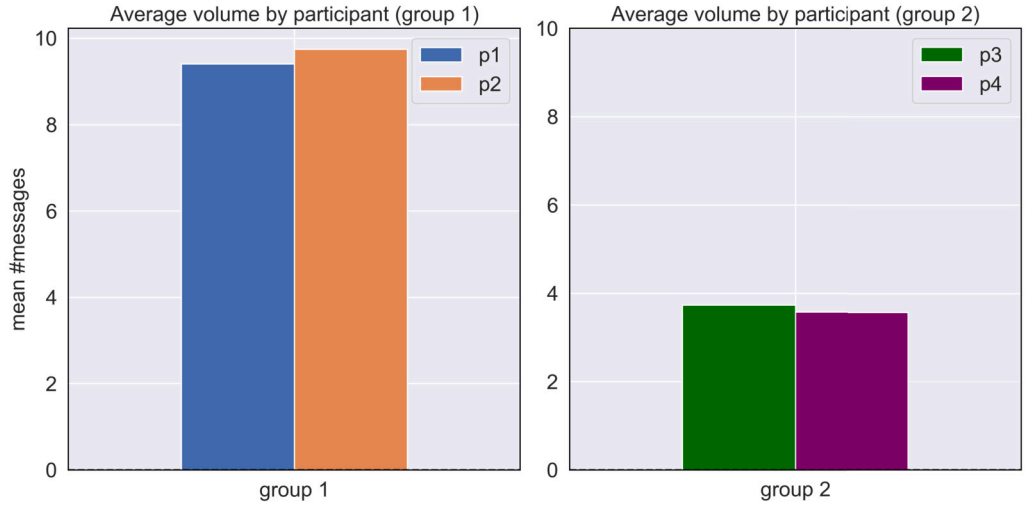

Fig. 2. Plots of the average volume for the whole conversation by group and participant, with group 1 (left) and group 2 (right).

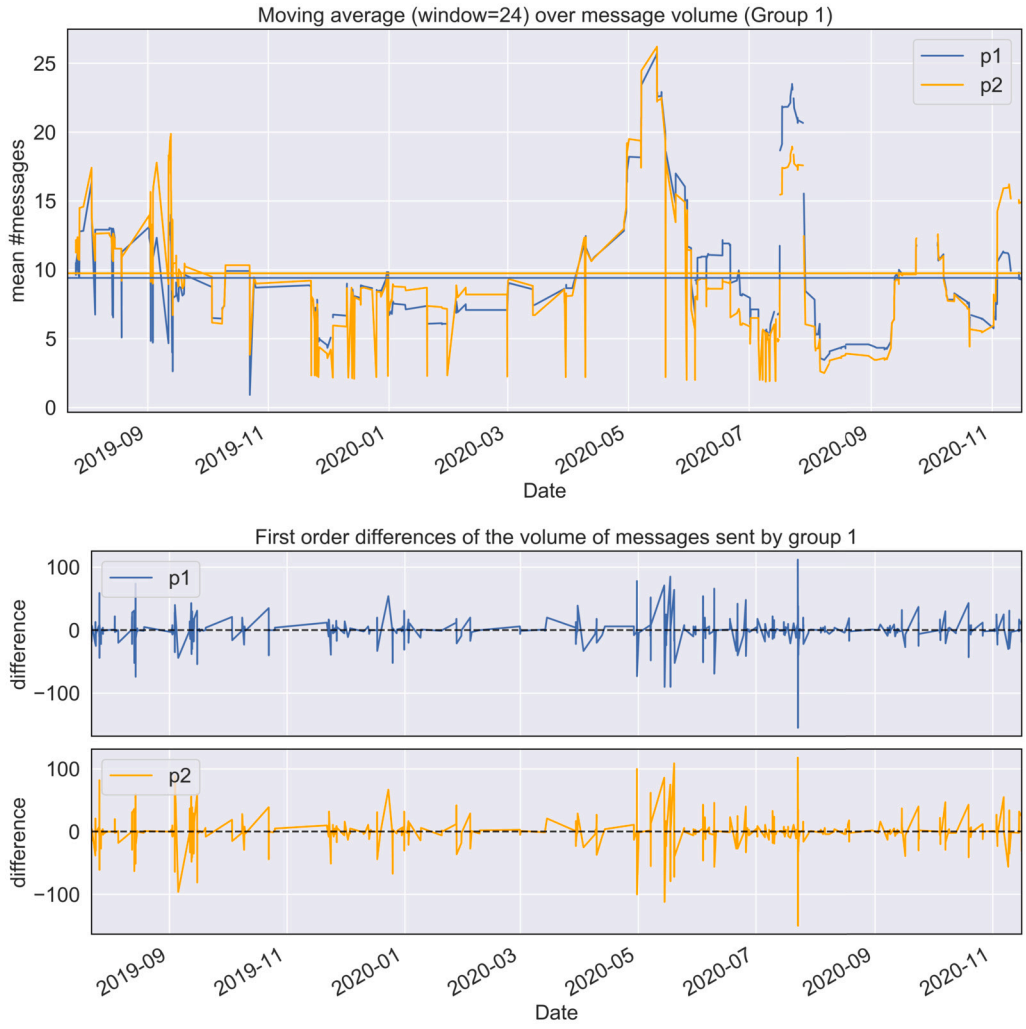

Fig. 3. The moving average (window = 24) for volume of messages sent for group 1 participants, with p_1, and p_2 (top). The horizontal lines represent the average volume of messages sent by each participant to act as a baseline. A further plot (bottom) shows the first order differences of the volume of messages sent for group 1.

4. We supplement the data used for training in the current run with the top-ten extracted sequences with high viterbi, and add them to the training data for the next run.
5. We load the trained model from the previous run, and use it as the starting point for the next run of training.
6. We then return to step 2, and repeat $k = 5$ times.

Fig. 4. The moving average (window = 168) for volume of messages sent for *group 1* participants (p_1 and p_2). The horizontal lines reflect the average volume of messages sent by participant to act as a baseline. A further plot (bottom) presents the first-order differences of the volume of messages sent for *group 1*.

After training we evaluate the model on a manually annotated sample from our data set of 200 sentences. When the accuracy for person and location exceeds 0.7 we proceed to label all the words in each message over the whole dataset, which we export together with the date and time of the message to convert them for use in the time series. In the next section, we discuss the results of obtained from the analysis of volume, sentiment, lexical diversity, and named entity recognition.

4. Results

In this section we present the time series generated from the message volumes, sentiment, lexical diversity, and named entities based on a moving average and first order differencing plots over a daily and weekly trend for each group and participant.

4.1. Time series analysis of message volume

We computed the average volume of messages sent for the whole period for each participant and group to create a baseline to identify when participant rate of communication increased or decreased suddenly and over a short period. The average volume per participant is presented in Fig. 2.

We first explore the daily trend (i.e., we set the moving-average window to 24 hours) for both groups with respect to their message volumes, with *group 1* presented in Fig. 3, and *group 2* presented in Fig. 5.

The time series over volume for *group 1* shows that there was regular communication between the two participants. On average each participant sent between 8 and 9 messages a day, denoted by the horizontal line for each participant. There were two periods with increased volume, one around January 2020, and another in June 2020. We observe that p_2 sent marginally more messages on average from the start of the period until February 2020, at which point the volume of messages sent by p_2 declined whereas the volume for p_1 were on average higher, particularly from June 2020. Furthermore, the time series highlights periods where there was less than average message volumes for p_1 at the start and extending to January 2020, followed by a further increase and again in June 2020. (See Fig. 4.)

In summary, for *group 1*, we noted that the moving average over volume revealed two potential events, one occurring in the lead up to January 2020, which resulted in p_1 sending marginally more messages to p_2 over the rest of the period. A second event occurring in June 2020, resulted in a general decline in the volume of communication over the last six months of the interaction after the largest and steepest increase in volume of messages exchanged over the whole conversation.

Turning to the moving average for *group2*, we observe that participants sent on average three messages per day (see Fig. 2). The lower average for *group1* may be indicative of a slightly different relationship. For instance, the participants may live together and engage largely in face-to-face communication. The moving average in Fig. 5 shows that there was an initial high volume of messages being sent at the start of

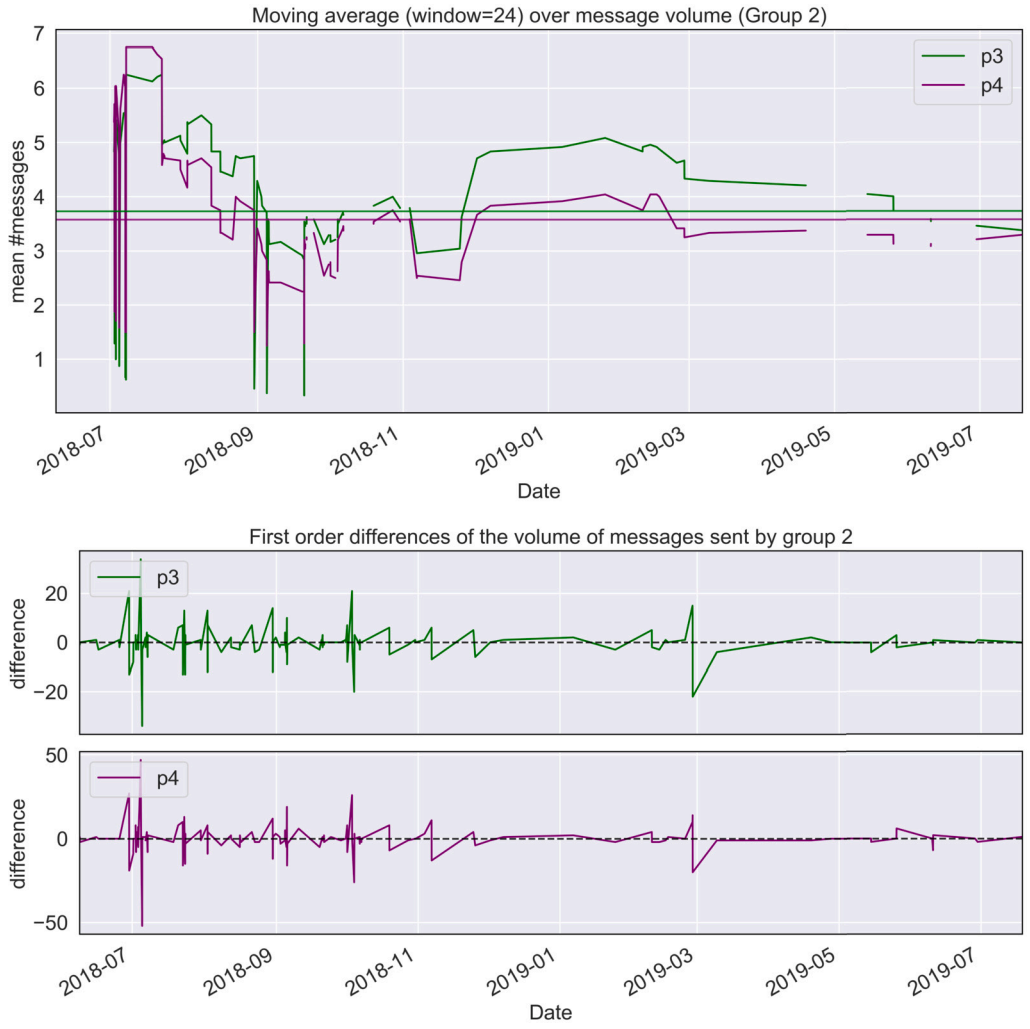

Fig. 5. The moving average (window = 24) for volume of messages sent by *group 2* participants (top). The horizontal lines reflect the average volume of messages sent by participant to act as a baseline. A further plot (bottom) presents the first order differences of the volume of messages sent for *group 2* participants (p_3 and p_4).

the period of communication, which was followed much lower volumes during the first month (July 2018). The volume reached a peak at the end of July 2018, and then began to steadily decrease over the course of three months to below average volumes for both participants.

After October 2018, we see the volume beginning to increase, and the distribution of messages sent over the period is fairly monotonic, but with p_3 sending more daily messages than p_4 over the last half of the period from November 2018.

We also noted breaks in communication lasting several weeks between May and July 2019. Looking at the difference plot (bottom of Fig. 5), we see that there were several points during communication with large differences in the number of messages sent daily at the beginning of the period followed by more stable periods in the volume of messages exchanged.

Note, there was no weekly trend generated for *group 2* due to a lack of data points. In summary, the daily trend in the volume of messages sent between participants in *group 2* (see top plot of Fig. 5), shows that p_3 sent on average more messages than p_4 throughout the interaction. The volume for both participants was at its highest point at the start of the interaction, whereupon it steadily decreased to approximately half the volume for the large part of the period. There were also breaks of several weeks towards the end, starting in the latter part of April 2019. The difference plot, bottom of plot of Fig. 5, shows periods of fluctuation in the volume of messages, with long periods of stability where participants are sending a constant volume of messages each day.

To conclude, the moving average over the volume of messages, reveals the trend and shows that the participants in each group communicated on a regular basis, but in lower volumes. The volume of exchanges between participants is similar between participants, with participants switching roles with respect to who is messaging at higher volumes, whereas p_3 generally sends marginally more messages to p_4, suggesting a leading role in the communication. In the next section we review the results of the sentiment analysis.

4.2. Time series analysis of message sentiment

In this section we present the results of the sentiment analysis applied to the message texts for each participant. We present several examples of the output of the sentiment analysis, and the results of a time series analysis based on moving averages produced on a daily and weekly basis together with the results of the first-order differentiation.

We first present a number of examples from the sentiment model including positive examples in Table 6, neutral in Table 7, and negative sentences presented in Table 8, respectively.

The moving average over sentiment, in Fig. 6 is the daily trend for *group 1* participants. The plot reveals that for *group 1* the sentiment between the two participants was marginally positive and rather stable until April 2020. At this point the sentiment for participant p_1 remained above their average sentiment, whereas the sentiment for p_2 began to decline and became increasingly negative, before falling below average

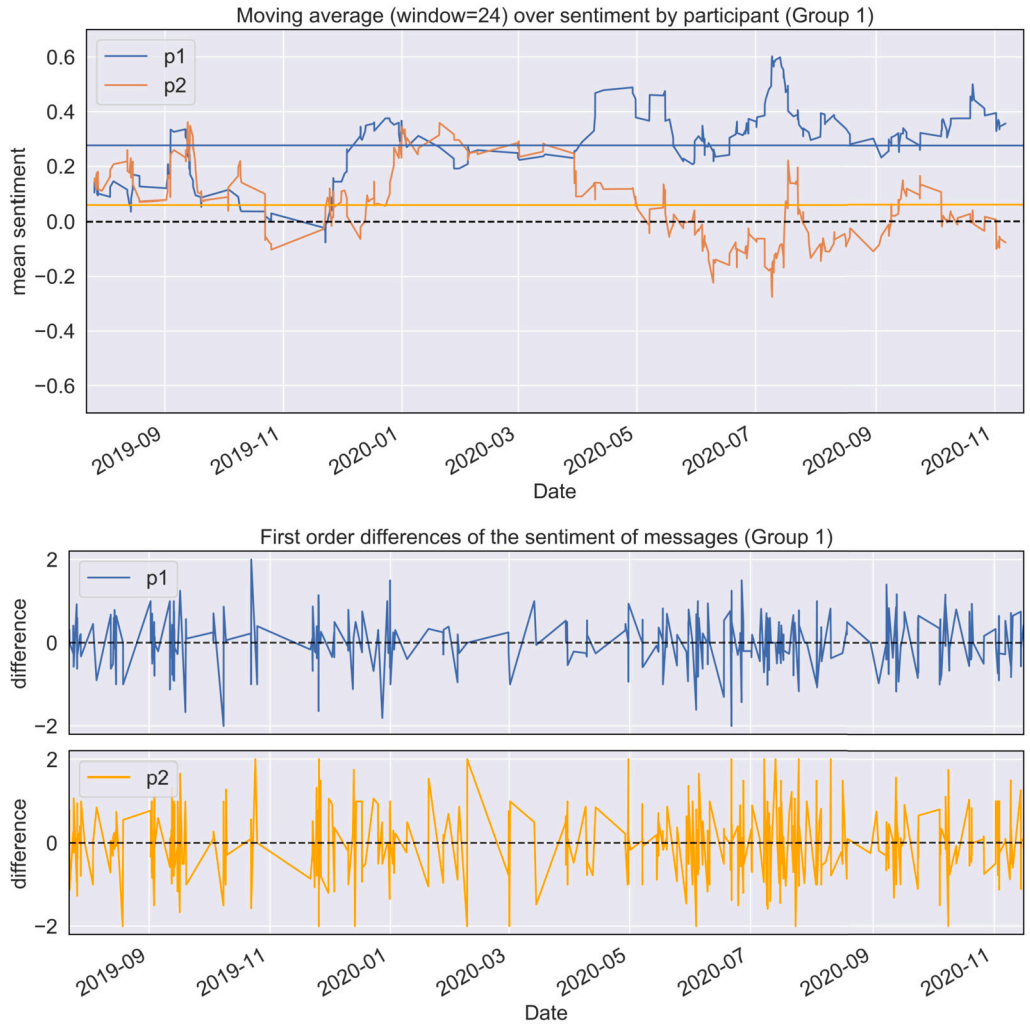

Fig. 6. The moving average (window = 24) for sentiment (top) of messages sent for group 1 participants (p_1 and p_2). The horizontal lines reflect the average sentiment of messages sent by participant to act as a baseline. We also present the first order differences of the sentiment (bottom) of the messages sent for *group1*.

Table 6
Positive example sentences returned by the sentiment classifier.

Message text	Score
Oh shittt wow	0.02
Oh yay awesome	0.14
Omg *J* with the "ovbs" yes useful contribution thank u sm	0.16
Ohh that's cool how's he finding it?	0.16
Aww that's so cute how long is he doing it for?	0.20
Today was good I was in a quieter part of the shop but still a lot of work	0.20

Table 7
Neutral example sentences returned by the sentiment classifier.

Message text	Score
Okayy that's fine then good timing. I'll wait at the bus stop	0.00
Hello what time are we meeting?	0.00
Ohh yeah fair	0.00
I'm at nero now	0.00
Are you okaayy?	0.00
And I'm in germany tomorrow !!	0.00

for p_2 over the remaining period. This is particularly clear when we look at the weekly trend for the sentiment in Fig. 8.

The daily trend in sentiment for *group 2*, presented in Fig. 7, revealed that p_3 began the interaction below their average sentiment for the period of conversation before becoming moderately positive, above their average, for the majority of the communication. Participant p_4 exhibited below average negative sentiment, which declined from the beginning of the interaction up until mid-September 2018 when it reached its lowest point. It then steadily increased and became relatively stable up until the end of the period.

Next we present the moving average and first-order differencing plots generated over the sentiment for *group 1* based on a weekly trend, with window = 168 hours, in Fig. 8. Note, that due to a lack of data points, it was not possible to generate the weekly trend for the sentiment time series for *group2*.

4.3. Time series analysis of lexical diversity

Here we present the results of the MATTR measure of Lexical Diversity (LD) applied to the messages of participants p_1 and p_2 from *group1*. Fig. 9 presents the moving average based (window = 24 hours) representing a daily trend in the MATTR score over the conversation period, with the horizontal lines reflecting the MATTR computed over a fixed window = 1000 by participant to act as a baseline to describe their average style.

A low MATTR score would suggest simple phrases, potentially with repetition, whereas as higher MATTR scores suggest that the language of the text messages is more verbose and with a higher ratio of unique words in the participants' lexicon.

The first thing we observe is that the lexical diversity starts high, but then steadily decreases over the period from September to November

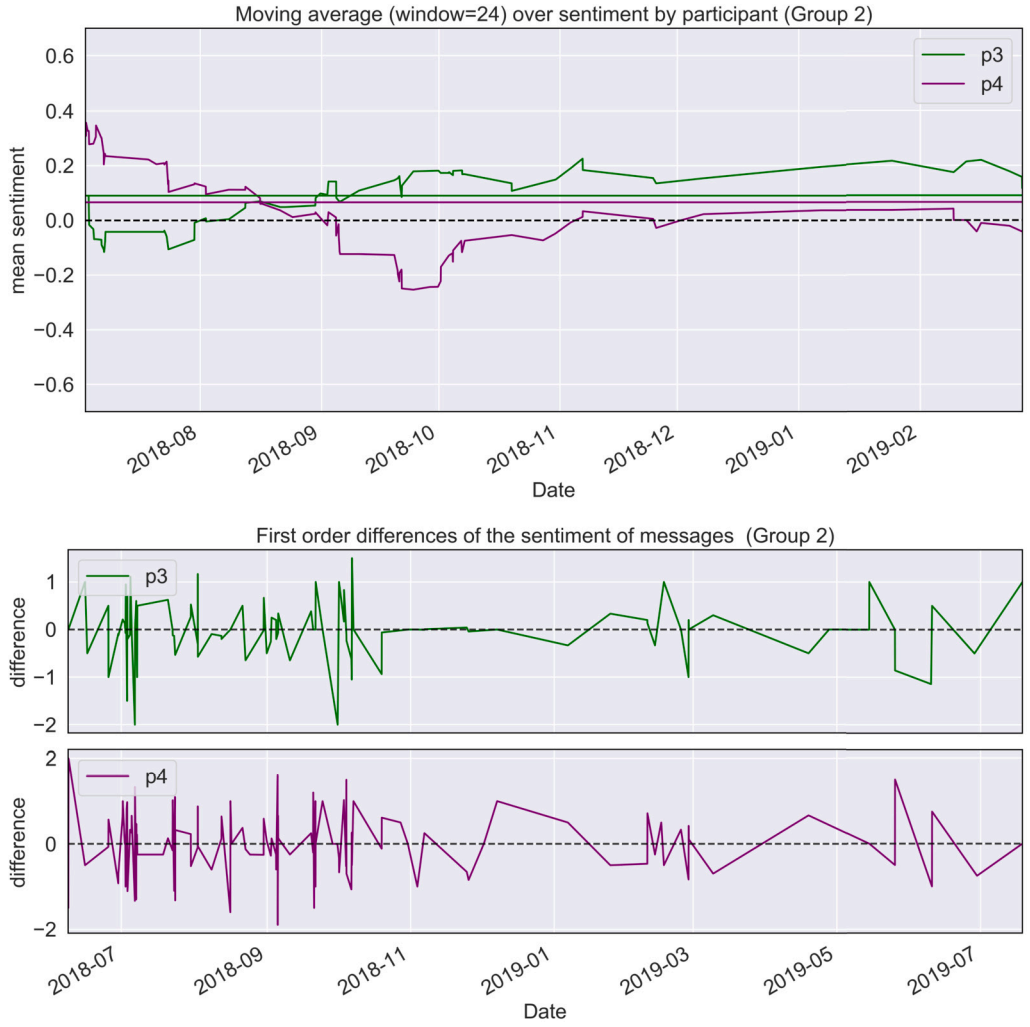

Fig. 7. The moving average (window = 24) for sentiment of messages (top) sent by group 2 participants (p_3 and p_4). The horizontal lines reflect the average sentiment of messages sent by participant to act as a baseline. We also present the plot of first order differences (bottom) for the sentiment of the messages.

Table 8
Negative example sentences returned by the sentiment classifier.

Message text	Score
Ffffffffs I'm already close to using up my data	−0.14
Those ppl are so loud omg	−0.18
Yeahh exactly so I'm sure she knows it would be awkward	−0.42
Yeah fuck I forgot cash tho. I'll try get some from my gran	−0.58
Yeahh it's annoying	−0.60
Ohh shit. Did her door get fucked?	−0.60

2019. After this from December 2019 to June 2020 we see that both participants are generally sending messages with a above average lexical diversity, with p_1 generally exhibiting higher MATTR scores, suggesting greater lexical diversity than p_2. Another point of interest is towards the end of the time series from July 2020 to October 2020, where both participants switch between above and below average MATTR scores in opposition to one another. That is, during the period, when the MATTR scores of p_1 increased, the MATTR scores of p_2 decreased in response, and vice-versa. This could provide an interesting starting point for investigators, since it suggests there is a potential event of interest e.g. an argument or debate.

When we look at the weekly trend of MATTR scores, we observe that the lexical diversity between participants is gradually declining over the period, with p_2 falling below their average MATTR. In summary, the moving average for the lexical richness reveals a similar pattern in that the MATTR measure for p_1 was marginally higher than p_2, whose MATTR score fell below average towards the end of the period, and the weekly trend shows that the MATTR for p_2 has been falling consistently over the period despite a small increase over the last month. (See Fig. 10.)

In Fig. 11, we observe a slightly different distribution in terms of lexical diversity, where both participants of *group 2* exhibit a similar time series "mirroring." The daily trend produced for the MATTR measure, in Fig. 11, shows that in general the messages of p_4 exhibited a greater lexical richness than p_3 throughout the period, increasing after the first month of interaction, reaching a peak in mid-September 2018, and then becoming stable for the remainder of the interaction after December 2012. In the last few weeks of the interaction, we see the MATTR scores of p_4 steadily drop, whilst p_3's increases in the same short period. This would warrant further investigation, since the MATTR scores are moving in opposite directions from the normal trend previously observed. This increase is also observed in the difference plot in Fig. 11. No plot was generated for *group 2,* due to a lack of data points to produce a weekly trend (window = 168 hours). In the next section, we present the results of the Named Entity Recognition.

4.4. Time series analysis of named entities

Here we present the results of the named entity recognition. Fig. 13, plots the volume of messages sent by each participant in *group 1* as

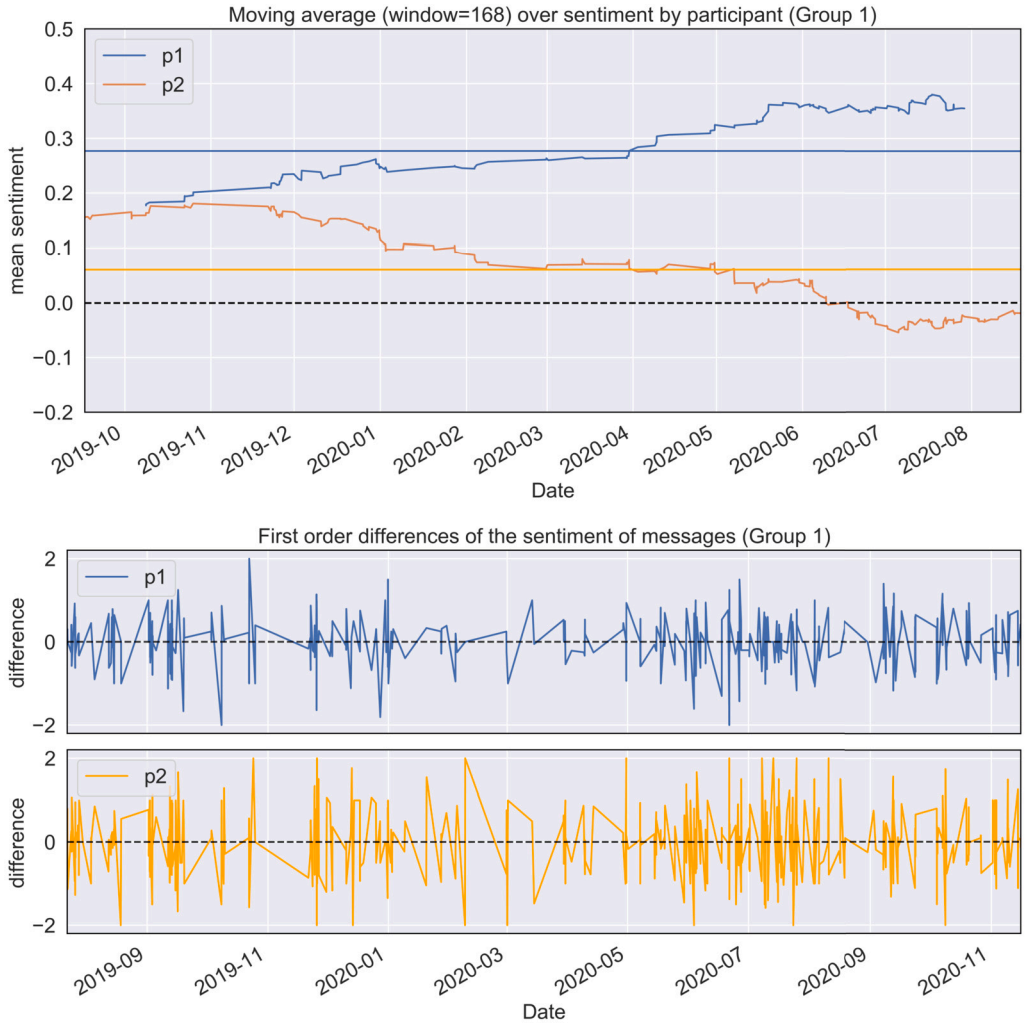

Fig. 8. The moving average (window = 168 hours) for sentiment of messages sent for group 1 participants, with p_1 (top), and p_2 (bottom). The horizontal lines reflect the average sentiment of messages sent by participant to act as a baseline. We also present the plot of first order differences (bottom) for the sentiment of the messages.

presented before, however, here we also overlay mentions of named entities related to location.

In the next example, in Fig. 12, the same technique can be applied to overlay all entities of a particular type, in this case all entities classed as a person name. Both approaches could be used to direct investigators to a subset of messages with mention of individuals. In the next example, we plot all entities of a particular type, in this case all entities classed as a person name. This demonstrates how the approach could be used to direct investigators to a subset of messages with mention of individuals.

The named entities can also be filtered to a specific entity. In this example, we filter the time series for location mentions of "London," represented as vertical lines illustrating when participants are discussing the location.

5. Conclusions and future work

We have presented a new framework designed to support the analysis of long-term personal conversations via SMS/WA which are typically extracted from mobile devices during an investigation. We did so in response to a set of informal requirements and desiderata that were elicited from British police and prosecutors whom were contacted as part of our externally-funded research project.

The first main contribution of our work is the design and validation of an architecture/dataflow customised for the practical scenario described by the Police and the CPS. At its core, there is text preparation and named-entity recognition by Samtla and semi-supervised sentiment analysis via pSenti. At the top, there is time-series visualisation and analysis, based on Conditional Random Fields.

The second result is the collection of preparation of a dataset with real data, which were instrumental in validating our ML operation. The results of our analysis of messages provided by *group 1* participants revealed two events in the daily trend in volume, a long-term decline in both sentiment and lexical richness as the communication between participants continued, which suggests the final event in June 2020 may have had an impact on their communication and potentially their relationship to cause a drop across the measures.[3]

Similar results were observed for *group 2*, 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

[3] Later on in the project, we manually inspected the conversations to check for information that would confirm that the dips in sentiment were related to actual events, e.g., a breakup or a discussion about a family member's welfare. So we could validate, albeit anecdotally, the output of our system.

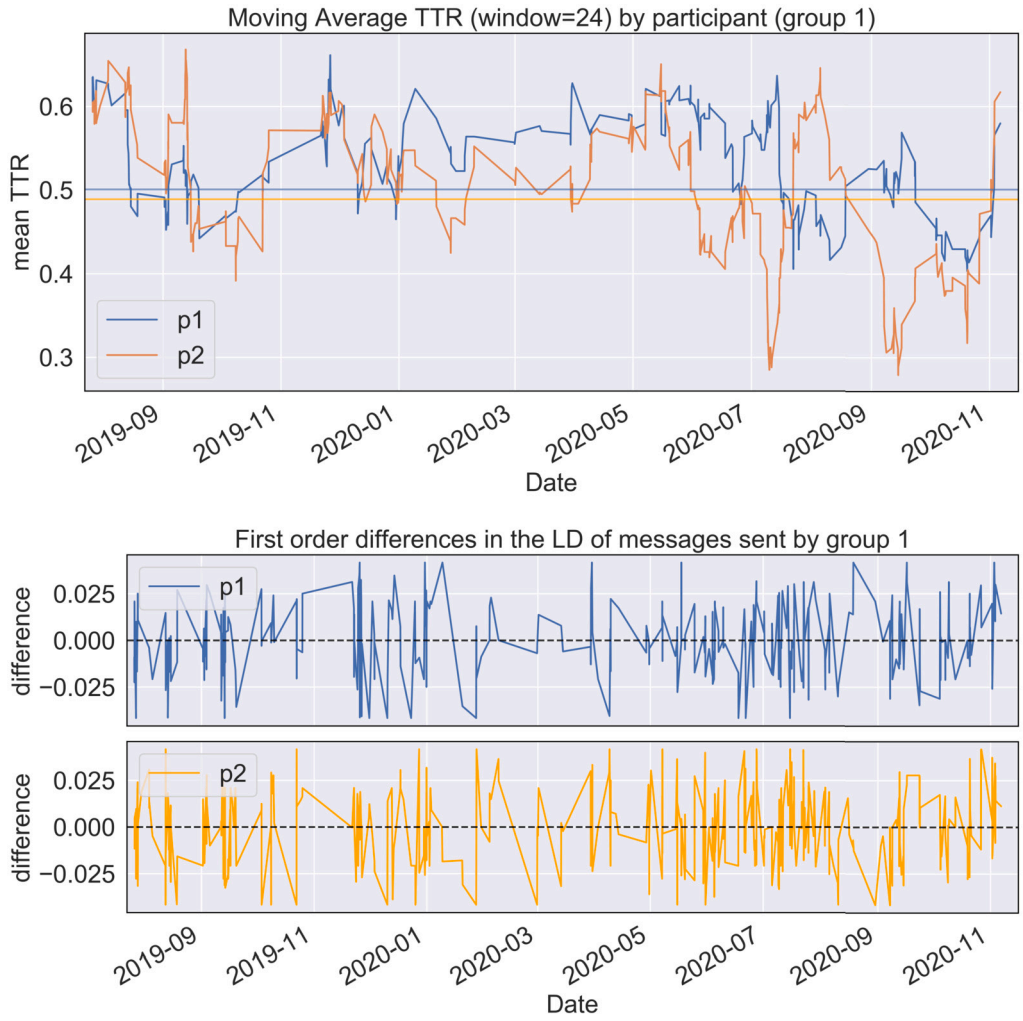

Fig. 9. The moving average (window = 24) for MATTR of messages sent for *group 1* participants, with p_1 (top), and p_2 (bottom). The horizontal lines reflect the MATTR computed over a fixed window = 1000 by participant to act as a baseline.

methods introduced, though short term trends may still provide some insight.

In terms of the sentiment analysis, it is important to ensure the quality of results in terms of accuracy of the sentiment and classification of named entities, which require domain specific gazetteers to be compiled by investigators, potentially from legacy data collected from previous cases. Investigators would then need to provide training and test examples to evaluate the quality of results.

The results of named entity recognition provides a means to detect novel terms such as drug names (identified by I-MISC entities), track movements of individuals over time through the location entities, and interactions with third parties and organisations through person and organisation entities.

Our framework is composed of a set of approaches that could be combined to aid the development of software tools and dashboard interfaces to aid investigators in a criminal investigation in the processing and analysis of digital information from mobile devices and social media platforms. The approaches require relatively little data, although we acknowledge that long-term trends are not always possible with small data sets, which may limit the analysis when conversations go back for years.

The volume, sentiment, 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 combination of approaches we introduced will aid in reducing the manual work in identify *regions* (time windows) of interest from a potentially-large data set. This is particularly the case when data spans long-term, with regular exchanges between multiple participants thus making it difficult for an investigator to gain quick insights into the interactions.

5.1. Direction for further research

Even though our focus has been on supporting investigators who look at SMS conversations to gather evidence of interest for criminal investigations, we believe that the lightweight architecture we presented here is of independent interest and could be developed in a number of ways.

Future work looks to develop methods in automated anomaly detection, based on machine learning techniques applied to time series data, which may reveal regions of interest for investigators to query in a semi-automatic way. The current framework could be implemented as part of a dashboard to enable investigators to explore their data and provide expert knowledge to retrain the sentiment classifier and NER models. We would like to deploy these models to real world data as part of a wider evaluation of the potential of these methods for aiding investigators in exploring digital texts composed of short social media messages.

We will also look at developing methods to identify the type of relationship between the participants and analyse how it evolves during

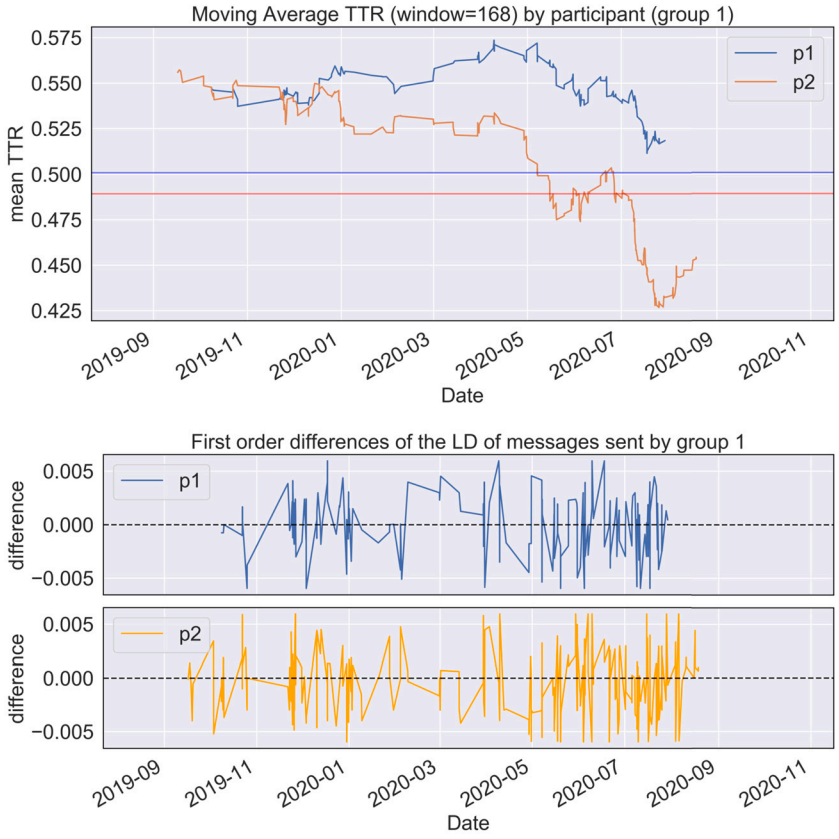

Fig. 10. The moving average (window = 168) for MATTR of messages sent for *group 1* participants, with p_1 (top), and p_2 (bottom). The horizontal lines reflect the MATTR computed over a fixed window = 1000 by participant to act as a baseline.

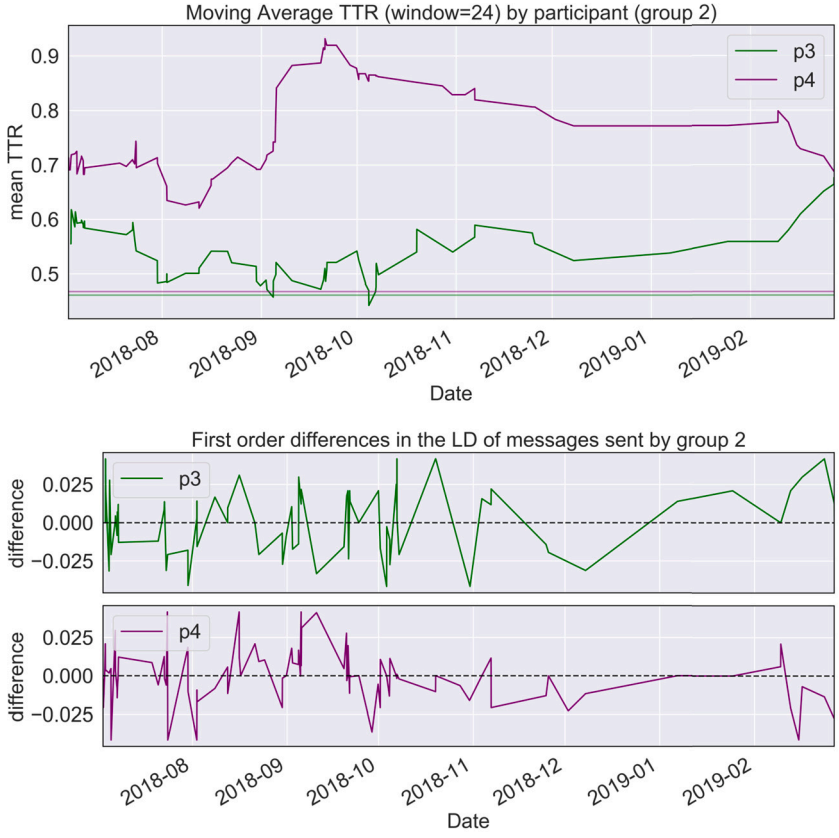

Fig. 11. The moving average (window = 24) for MATTR of messages sent for *group 2* participants, with p_3 (top), and p_4 (bottom). The horizontal lines reflect the MATTR computed over a fixed window = 500 by participant to act as a baseline.

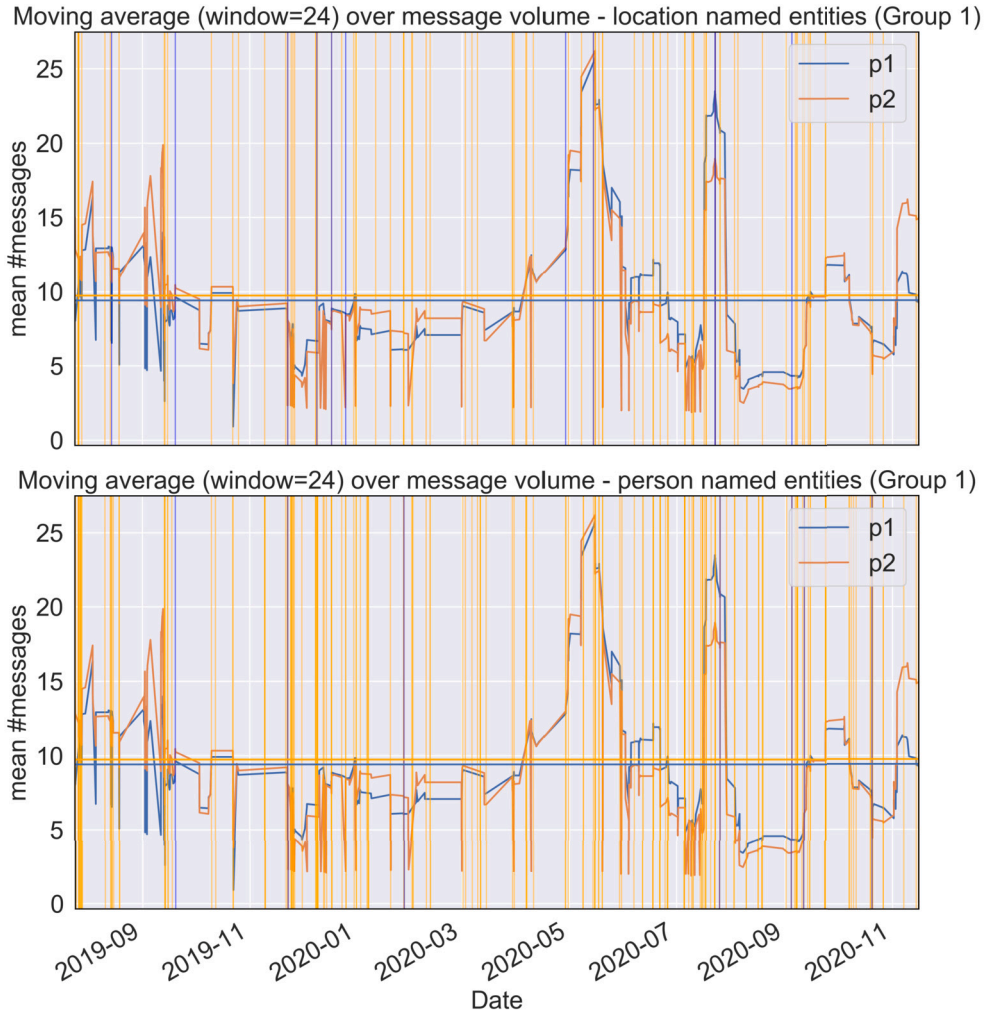

Fig. 12. The moving average with (window = 24 hours) representing the daily trend for *group 1. Vertical lines here represent mentions of people by both participants.*

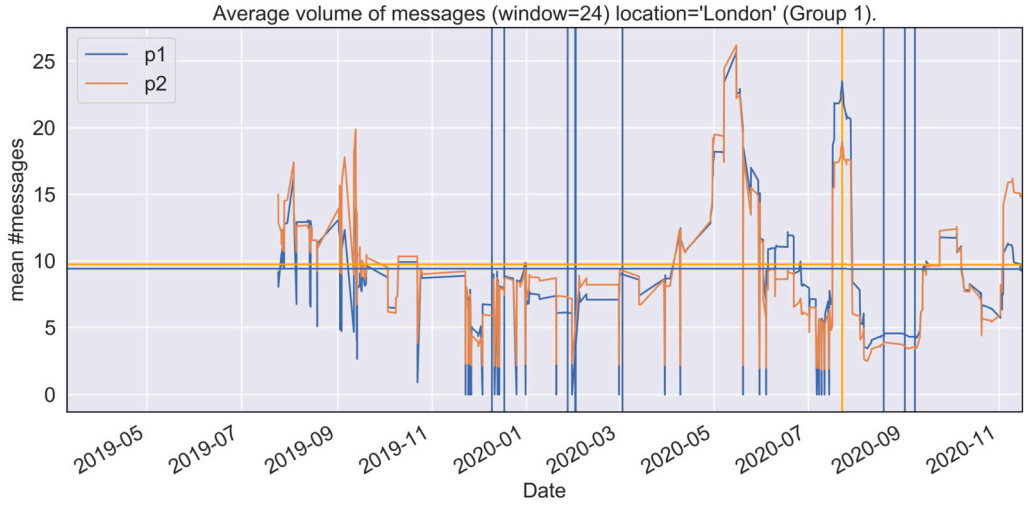

Fig. 13. The moving average with (window = 24 hours) representing the daily trend for *group 1. Vertical lines represent mentions of the named entity location 'London' by both participants.*

the conversation. With enough data real world and case studies, it may be possible to construct a series of profiles, including an friendship, an intimate relationship, abusive relationship, and logistical conversations involving organised crime.

CRediT authorship contribution statement

Martyn Harris: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Vi-

sualization, Writing – original draft, Writing – review & editing. **Jessica Jacobson:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision. **Alessandro Provetti:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

We are in debt to the four anonymous volunteers who made this research possible by donating their most private conversations.

This research was made possible thanks to the generous funding provided by the Dawes Trust to the project “Digital forensics and social media: Challenges and opportunities for law enforcement” led by Prof. J. Jacobson in 2019-2023.

References

Casey, E., 2019. The chequered past and risky future of digital forensics. Aust. J. Forensic Sci. 51, 1–16. https://doi.org/10.1080/00450618.2018.1554090.

Chatfield, C., 2004. The Analysis of Time Series: An Introduction. CRC Press, Florida, United States.

Covington, M., McFall, J., 2010. Cutting the Gordian knot: the moving-average type-token ratio (MATTR). J. Quant. Linguist. 17, 94–100. https://doi.org/10.1080/09296171003643098.

Coyac-Torres, J.E., Sidorov, G., Aguirre-Anaya, E., Hernández-Oregón, G., 2023. Cyber-attack detection in social network messages based on convolutional neural networks and NLP techniques. Mach. Learn. Knowl. Extr. 5, 1132–1148. https://doi.org/10.3390/make5030058. https://www.mdpi.com/2504-4990/5/3/58.

Fergadiotis, G., Wright, H., West, T., 2013. Measuring lexical diversity in narrative discourse of people with aphasia. Am. J. Speech-Lang. Pathol. 22, S397–S408. https://doi.org/10.1044/1058-0360(2013/12-0083).

Harris, M., Levene, M., 2021. SamtlaAPI: free your data. Online. http://www.samtla.com/. (Accessed 15 April 2024).

Harris, M., Levene, M., Mudinas, A., 2024. Time series analysis of sentiment: a comparison of the US and UK coronavirus subreddits. Int. J. Inf. Technol. Decis. Mak. 23, 57–88. https://doi.org/10.1142/S0219622023400035.

Holt, T., Bossler, A., Seigfried-Spellar, K., 2015. Cybercrime and digital forensics: an introduction. https://doi.org/10.4324/9781315296975.

Hussein, D.D., 2018. A survey on sentiment analysis challenges. J. King Saud Univ., Eng. Sci. 30, 330–338. https://doi.org/10.1016/j.jksues.2016.04.002. https://www.sciencedirect.com/science/article/pii/S1018363916300071.

Krishnan, S., Shashidhar, N., Varol, C., Islam, A., 2022. Sentiment analysis of case suspects in digital forensics and legal analytics. Int. J. Secur. 13.

Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

Liu, B., 2015. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Mudinas, A., Zhang, D., Levene, M., 2012. Combining lexicon and learning based approaches for concept-level sentiment analysis. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM ‘12), vol. 5. Association for Computing Machinery, New York, NY, USA, pp. 1–8.

Mudinas, A., Zhang, D., Levene, M., 2018. Bootstrap domain-specific sentiment classifiers from unlabeled corpora. Trans. Assoc. Comput. Linguist. 6, 269–285. https://doi.org/10.1162/tacl_a_00020.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvist. Investig. 30, 3–26.

Studiawan, H., Sohel, F., Payne, C., 2020. Sentiment analysis in a forensic timeline with deep learning. IEEE Access 8, 60664–60675. https://doi.org/10.1109/ACCESS.2020.2983435.

Sutton, C., McCallum, A., 2012. An Introduction to Conditional Random Fields. Now Publishers Inc., Hanover, MA, USA.

Torruella, J., Capsada, R., 2013. Lexical statistics and tipological structures: a measure of lexical richness. In: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013). Proc., Soc. Behav. Sci. 95, 447–454. https://doi.org/10.1016/j.sbspro.2013.10.668.

Tully, G., Cohen, N., Compton, D., Davies, G., Isbell, R., Watson, T., 2020. Quality standards for digital forensics: learning from experience in England and Wales. Forensic Sci. Int. 32, 200905. https://doi.org/10.1016/j.fsidi.2020.200905.