Popular Detectives Stories Writing Styles: Unchanging Overall Sentiment and Dominant Cognitive Mode

Yaning Shi (1004410559) *
University of Toronto

April 8, 2022

Abstract

Detective stories originated in mid-19th century and Mystery/Crime has become one of the most popular genres in fiction. To explore some common features in popular detective stories, information was extracted from detective stories with full text access on the internet. Using methods of regression and topic modelling, results from statistical analysis show that 1)detective stories involving murder and trafficking have longer portion of text on crime-solving process; 2)sentiment of a detective story tends to maintain on a positive-negative scale; 3)content of a detective storie usually reflect a dominant cognitive mode throughout.

Keywords: Crimes, Lexical semantics, Sentimental analysis, Semisupervised Latent Dirichlet allocation (seeded-LDA), Topic modelling

*Applied statistics student with focus in health studies. Special thanks to project collaborators: Adam Hammond (Department of English), Simon Stern (Faculty of Law), TA advisor: Marjia Pejcinovska (Department of Statistical Sciences)

1 Introduction

This report describes a few statistical analyses on some popular detective stories and concluded common features as results, and is a part of "the Birth of the Modern Detective Story" project.

The project was initiated by two professors at the University of Toronto: Adam Hammond from Department of English and Simon Stern from Faculty of Law. They are interested in finding some common features in popular detective stories and hired a group of annotators to read detective stories with full text readily available on the internet and extract information from each story. Data was collected on 346 popular detective stories written by 18 authors and are provided along with corresponding plain texts. The extracted data includes authors' demographic information, summaries of people and events involved in the stories, and description of the narratives.

A regular detective story is composed of two parts: investigation and reveal, divided by a border sentence. In this report, we aim to address the following research topics:

- Relationship between types of crimes involved and length of reveal
- Sentiment difference in before and after the investigation-reveal border
- Cognition difference in before and after the investigation-reveal border

After this introduction the rest of the report is comprised of the following: description of selection and processing of data in section 2, illustration of statistical methods used in each of the research topics in Section 3, presentation of data analysis outcomes in Section **??**, results interpretation and comments on limitations in Section 4. Technical figures for optional reading will be placed in the supplementary material in Section 5.

2 Data Summary

Only regular detective stories with a border sentence that divides investigation and reveal are used in the analysis, each border sentence is included in either investigation or reveal, whichever the story text ends with. Uninterpreted non-metacharacter escapes, punctuation, symbols, and separators are removed from story texts and border sentences.

Investigation and reveal lengths of story texts are measured in their number of characters. Relative reveal length is the ratio between reveal and investigation lengths of the text.

A set of binary variables are used to indicate the presence of different type of (quasi-)crimes in the detective stories, the (quasi-)crimes include: suspected murder, theft, fraud, mischief, assault, murder, forgery, kidnapping, breaking and entering, blackmail, trafficking, bribery, and illegal gambling.

3 Methods

3.1 Relationship between length of reveal and types of (quasi-)crimes involved

The aim is to determine the effects of different types of (quasi-)crimes have on the relative length of reveal while accounting for effects associated with individual authors due to personal writing styles, which include preferred types of (quasi-)crimes in their writings. The relative lengths of reveal are ratios between reveal and investigation lengths of detective story text, they range from 0.03 to 6.55 and the distribution skews to the right. We therefore assume a gamma distribution for relative reveal length, for that gamma distribution describes a continuous probability on the positive axis with two separate parameters which together allow a flexible shape.

We fit a generalized linear mixed model (GLMM) between the relative lengths of reveal and the types of (quasi-)crimes involved, in which the random effects on the author level are accounted for on the basis of a generalized linear model (GLM). A GLM works just like an ordinary linear model, and can also be used for response data that do not follow a normal distribution. The difference in GLM that allows this extended application is a link function that is applied to the non-normal data so that they are normalized. We use log as the link function to normalize the relative reveal lengths variable. Final model is chosen by minimizing AIC in a stepwise algorithm. More details on data normalization (Section 5.1.1), model specifications (Section 5.1.2), model selection (Section 5.1.3), multicollinearity and residual verification (Section 5.1.4) are discussed in the supplementary material for this

report.

3.2 Sentiment difference in before and after the investigation-reveal border

To identify sentiment patterns between portions before and after the investigation-reveal border of detective stories, we first use topic modelling to identify the overall sentiment of each text segment. Create a topic model with the sentiment-defining elements of the 2015 Lexicoder Sentiment Dictionary (Young & Soroka 2012) as seed tokens (words and phrases). The dictionary defines sentiments on a positive-negative scale on 4 different levels: negative, negative-positive, negative-negative, and positive. The negative-positive and negative-negative sentiments are defined with phrases with positive/negative words preceded by a negation. The full dictionary is attached in the supplementary material (Section 5.4). Implement semisupervised Latent Dirichlet allocation (seeded-LDA) (Lu et al. 2011) over a pool of tokenized detective story texts to aquire the new topic model adjusted for use in the context of detective stories. Model specifications can be found in the supplementary material (Section 5.2.1). Use the new topic model to categorize each detective story text segment of investigation or reveal for its overall sentiment. A preview of the new LDA model attached in the supplementary material (Section 5.2.2) gives a general idea of the model "fit".

Next, identify any significant sentiment patterns between portions of detective stories before and after investigation-reveal border with Pearson's Chi-squared Test for Count Data and post hoc analysis based on the standardized residuals (Beasley & Schumacker 1995). The null hypothesis for the Pearson's Chi-squared Test assumes independence of overall sentiment in before and after investigation-reveal border, and that for the post hoc analysis assumes independence of the two categorical variables on individual levels. The p-values from the post hoc analysis are then adjusted with Holm's correction method (Holm 1979) to solve multiplicity from the multiple comparison procedures by controlling family-wise error rate to be less than equal to our significance level 0.05. Procedure of the post hoc analysis and algorithm of Holm method are shown in Section 5.2.3 in the supplementary material. A p-value from Pearson's Chi-squared Test lower than 0.05 would suggest rejection of the

null hypothesis for Pearson's Chi-squared Test and that overall sentiments in before and after investigation-reveal border are not independent; an adjusted p-value from the post hoc analysis lower than significance level would suggest that the corresponding residual contributes to rejection of null hypothesis in Pearson's Chi-squared Test.

3.3 Cognition difference in before and after the investigation-reveal border

We identify cognition patterns between portions before and after the investigation-reveal border of detective stories using an approach similar to that used for identifying sentiment differences. Different than method described in Section 3.2, the initial topic model is created with Collin Martindale's Regressive Imagery Dictionary Martindale (1975) Martindale (1990). The dictionary defines cognition on a primordial-conceptual scale on 3 different levels: primary process cognition, emotions, and secondary process cognition. Primary process cognition is considered primordial and its content includes drives, perceptual disinhibition, sensation, regressive cognition, and Icarian imagery. Secondary process cognition is considered conceptual and its content includes temporal references, moral imperatives, instrumental behavior, social behavior, abstraction, restraint and order. Emotions are neither primordial or conceptual and are in on the middle of the scale. The full dictionary (Section 5.4) and a preview of the new LDA model (Section 5.3.1) are attached in the supplementary material.

We identify any significant cognition patterns between portions of detective stories before and after investigation-reveal border with Pearson's Chi-squared Test for Count Data and post hoc analysis based on the standardized residuals (Beasley & Schumacker 1995). The null hypothesis for the Pearson's Chi-squared Test assumes independence of overall sentiment in before and after investigation-reveal border, and that for the post hoc analysis assumes independence of the two categorical variables on individual levels. The p-values from the post hoc analysis are then adjusted with Holm's correction method (Holm 1979) to solve multiplicity from the multiple comparison procedures by controlling family-wise error rate to be less than equal to our significance level 0.05. Procedure of the post hoc analysis and algorithm of Holm method are shown in Section 5.2.3 in the supplementary material.

A p-value from Pearson's Chi-squared Test lower than 0.05 would suggest rejection of the null hypothesis for Pearson's Chi-squared Test and that overall sentiments in before and after investigation-reveal border are not independent; an adjusted p-value from the post hoc analysis lower than significance level would suggest that the corresponding residual contributes to rejection of null hypothesis in Pearson's Chi-squared Test. # Results ## Relationship between length of reveal and types of (quasi-)crimes involved Summary of the final GLMM (Table 1) and Table 2 shows statistically significant relationship between relative length of reveal and some types of (quasi-)crimes. Detective stories being written by different authors explains $\frac{0.1956}{(0.1956+0.8035)} \times 100\% = 19.47\%$ of the total variance in relative reveal length. On a 0.05 significance level, irrelevant of being written by different authors, adding suspected murder, kidnapping or trafficking to the plot of a detective story without changing the presence of other (quasi-)crimes has an effect on relative reveal length, with log(relative reveal length) decreases by 0.3849, increases by 0.2924 and decreases by 1.2630 with the presence of suspected murder, kidnapping, and trafficking, respectively.

Table 1: Summary of Random Effects of the Final GLMM

grp	var1	var2	vcov	sdcor
author_code	(Intercept)	NA	0.1956	0.4423
Residual	NA	NA	0.8035	0.8964

Table 2: Fixed Effects of the Final GLMM

	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-0.7556	0.1488	-5.0772	0.0000
involves_suspected_murderTRUE	-0.3849	0.1521	-2.5310	0.0114
involves_murderTRUE	-0.1538	0.0997	-1.5429	0.1229
involves_fraudTRUE	-0.0822	0.0899	-0.9149	0.3602
involves_kidnappingTRUE	0.2924	0.1357	2.1552	0.0311

	Estimate	Std. Error	t value	Pr(>|z|)
involves_traffickingTRUE	-1.2630	0.3532	-3.5754	0.0003

3.4 Sentiment difference between before and after the investigation-reveal border

Pearson's Chi-squared Test for Count Data gives a p-value of $1.272e^{-12}$. Therefore on a 0.05 significance level, we reject the null hypothesis and accept that the overall sentiment before the investigation-reveal border is correlated with that after. Table 3 shows standardized residuals and p-values from post hoc analysis. On a 0.05 significance level we conclude that these combinations of overall sentiment of first part and second part are positively correlated:(negative, negative), (negative-positive, negative-positive), (negative-negative, negative-negative); and the combination (negative-positive, negative) is negatively correlated.

Table 3: Standardized Residuals and P-Values from the Post Hoc Analysis

Dimension	Value	negative	negative_positive	negative_negative	positive
negative	Residuals	4.1823	-2.2543	-2.1265	-0.4877
negative	p values	0.0005	0.3626	0.4685	1.0000
negative_positive	Residuals	-3.0505	6.2988	-0.5083	-2.3902
negative_positive	p values	0.0343	0.0000	1.0000	0.2358
negative_negative	Residuals	-2.7133	-2.0604	4.9681	0.3424
negative_negative	p values	0.0999	0.5510	0.0000	1.0000
positive	Residuals	1.0868	-1.5565	-2.2317	2.5778
positive	p values	1.0000	1.0000	0.3846	0.1591

3.5 Cognition difference in before and after the investigation-reveal border

Pearson's Chi-squared Test for Count Data gives a p-value of $1.404e^{-70}$. Therefore on a 0.05 significance level, we reject the null hypothesis and accept that the overall cognitive mode before the investigation-reveal border is correlated with that after. Table 4 shows standardized residuals and p-values from post hoc analysis. On a 0.05 significance level we conclude that all combinations of the same overall cognitive mode in first part and second part are positively correlated; and all combinations of different overall cognitive modes are negatively correlated.

Table 4: Standardized Residuals and P-Values from the Post Hoc Analysis

Dimension	Value	primary	emotions	secondary
primary	Residuals	11.5977	-5.3935	-6.4861
primary	p values	0.0000	0.0000	0.0000
emotions	Residuals	-4.7631	12.5142	-7.3274
emotions	p values	0.0000	0.0000	0.0000
secondary	Residuals	-6.9381	-7.4192	14.2076
secondary	p values	0.0000	0.0000	0.0000

4 Conclusion/Discussion

4.1 Relationship between length of reveal and types of (quasi-)crimes involved

The results implies that detective stories with presence of murder or trafficking generally have a relatively shorter reveal than that without; detective stories with presence of kidnapping generally have a relatively longer reveal than that without. A possible reason for this could be that murder and trafficking crimes are difficult to solve and kidnapping crimes

are easy to solve, but the story doesn't end at the point where crime is solve, the process of tracking the kidnapper down is also in the story plot.

The GLMM does not explain a very large portion of the variance, this can be improved by adding more variables to the model with caution, because variables of different scales can be influential to variance composition.

4.2 Sentiment/Cognition difference between before and after the investigation-reveal border

Three of the sentiment levels in first parts of the stories are positively correlated with the same sentiment levels in second parts of the stories and one sentiment level is negatively correlated with a sentiment level on the opposite half of the scale; All of the cognitive modes are positively correlated with the same mode and negatively correlated with the other modes in the two parts. This implies that a popular detective story is likely to have a uniform sentiment on the positive-negative scale and the same dominant cognitive mode throughout the texts. There are two factors in the writing of a detective story that could affect the overall sentiment and dominant cognitive mode: story plot and narrative. A possible explanation is that a popular detective story could usually has a unchanging theme and a "narrator" (the author or a fictional narrator) that has a stable state of mind (not necessarily healthy, but a fixed mentality). Conversely, another explanation would be that story plot with a changing theme could usually be neutralized with a changing narrative, as how a change in internal thoughts could affect one's perception of objective reality (realization as an example); a story with changing narrative could usually be neutralized with red herrings (things that are later forced to be unlearned), as how external reality can force a change in interal thoughts. In either case, especially the the second one, it can be inferred that a uniform writing style is preferred in a story, and instability in either story plot or narrative is usually compensated.

Each detective story is segmented into 2 pieces and categorization is based on 3-4 levels, the results provides some intuition but frequent and subtle changes might not be captured. A possible way to resolve this is to segment each story into multiple pieces and use categorization scale with finer gradient or a discrete set of a larger number of categories.

Categorization of text segment with seeded-LDA method is sensitive to trimming of tokens, trimming schemes need to be made on a data-specific manner. Trimming of tokens is essential to avoid dominance in one category when the text segments have similar compositions which makes the categorization redundant. It also cuts down the model fitting runtime which is necessarily long due to large amount of documents and lengthy text segments. Excessive trimming can lead to failure of categorization, for that the likelihood scores of a text segment could be equal over all the categories from an oversimplified topic model.

5 Supplementary Material

5.1 Relationship between length of reveal and types of (quasi-)crimes involved

5.1.1 Data normalization

Quantile-quantile plots (Figure 1) show that relative reveal length is normalized by a log transformation, use "log" as the link function instead of the default "inverse" to optimize the model fit:

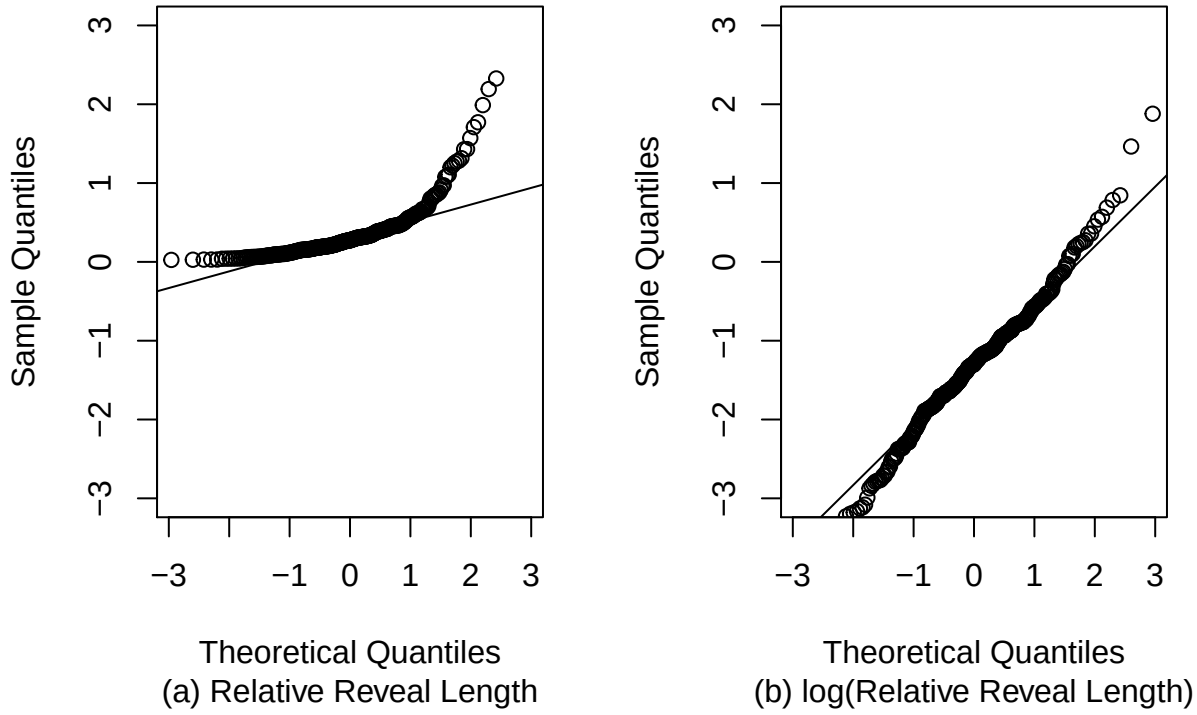

Figure 1: Normalizing relative length of reveal by log function: two Normal Q-Q Plots

5.1.2 Model specifications

We ssume the relative lengths of reveal has a gamma distribution:

$$Y_{ij}|U_i \sim Gamma(\mu_{ij}/\nu, \nu)$$

where Y_{ij} stands for a measurement of relative reveal length, μ_{ij}/ν the range parameter and ν the shape parameter. Fit a GLMM between relative lengths of reveal and types of

(quasi-)crimes involved, with log as the link function:

$$log(\mu_{ij}) = X_{ij}\beta + U_i$$

$$U_i \sim MVN(0, \Sigma)$$

where X_i is a vector of binary variables indicating presence of each type of (quasi-)crime in a story, $U_i; i = 1, \ldots, M = 18$ the random effects on the author level following a multivariate normal distribution with location 0 and covariance Σ.

5.1.3 Model selection

Fit a GLM with the fixed effect components of the proposed GLMM. Components of the model are searched stepwise and backwards for lowest Akaike Information Criterion (AIC) in nested models. The fixed effects of the selected model are listed in the Table 5 below:

Table 5: Fixed Effects of the Selected GLM

	x
(Intercept)	-0.7787
involves_suspected_murderTRUE	-0.2692
involves_fraudTRUE	-0.1592
involves_murderTRUE	-0.2402
involves_kidnappingTRUE	0.4280
involves_traffickingTRUE	-0.9162

Substitute the fixed effects part of the formula of the initially proposed GLMM with the selected GLM formula to get the final GLMM.

5.1.4 Assumptions verification

Variance Inflation Factors for fixed-effect variables in the final model are shown in Table 6:

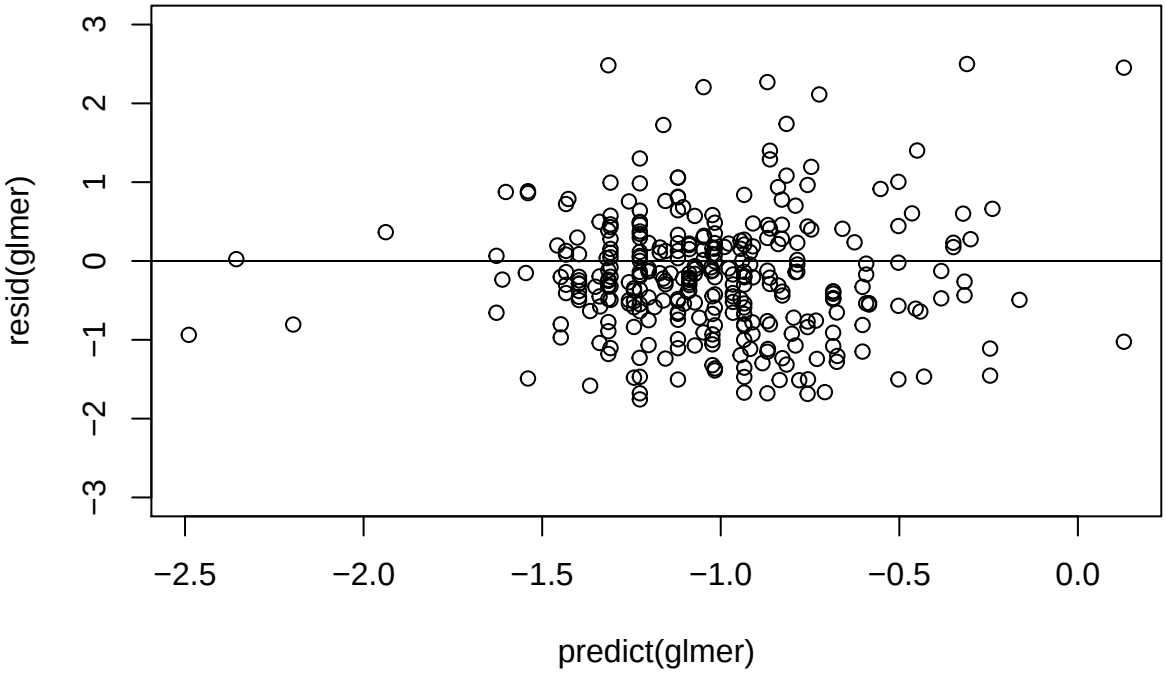

Figure 2: Residual Plot of the GLMM

Table 6: Variance Inflation Factors for the Fixed-Effect Variables in Final GLMM

	x
involves_suspected_murder	1.2029
involves_murder	1.2597
involves_fraud	1.0696
involves_kidnapping	1.0206
involves_trafficking	1.0143

All values are between 1 and 2, no special attention for multicollinearity is needed. Residuals from the final model are plotted against their predicted values in Figure 2 :

The points concentrate around x-axis and show symmetry.

5.2 Sentiment difference between before and after the investigation-reveal border

5.2.1 Model Specifications

Use the first topic model from seed tokens as the prior distribution for topics:

$$\phi \sim Dir(\{\beta + C_w\}_{w \in V})$$

where $\beta = 0.1$ is the topic-token probability as asymmetric smoothing prior, $C_w = 0.01$ is the equivalent pseudo counts added to sufficient statistics of the topic where each seed token w belongs as a asymmetric prior. The prior distribution for text segments is:

$$\theta \sim Dir(\alpha)$$

where α =50/(number of topics) is the document-token as a smoothing prior. For each token w_i (compounded with the 2015 Lexicoder Sentiment Dictionary and Wordnet Based Categorization Dictionary(Fellbaum 2006), with stopwords from Snowball(Porter 2001), frequently appeared tokens (top 50%), and tokens appearing in more that 90% of the text segments trimmed off) in text segment a topic is chosen:

$$z_i \sim Multinomial(\theta)$$

and a token is chosen:

$$w_i \sim \phi_{z_i}.$$

5.2.2 Model Fit

A preview of the new LDA model components is shown as Table 7:

Table 7: Preview of the New Sentiment Topic Dictionary

negative	positive	neg_positive	neg_negative
cringed	clears	not_beautiful	not_unlikely
pinching	cheeriness	not_extraordinary	not_unnaturally

negative	positive	neg_positive	neg_negative
monotony	glistened	not_satisfy	not_offended
estranged	fearless	not_responsible	not_hard
disagreeably	absolved	not_cleared	not_ghost
affectation	humours	not_connect	not_worse
invaders	exhilarating	not_encourage	not_shake
indisposition	sobered	not_accepted	not_bite
dishonourable	open-eyed	not_discovered	not_dark
repelled	complimenting	not_obliged	not_disappoint
egotism	aristocrat	not_benefit	not_nervous
shortcomings	sheltering	not_interested	not_bother
saddest	firmness	not_afford	not_steal
sulking	civilian	not_daring	not_stolen
insensibility	smart-looking	not_convincing	not_fallen
famished	sociable	not_free	not_scrap
sedentary	tolerable	not_apt	not_blind
provoked	keen-faced	not_learn	not_exaggerated
dark-lantern	achievements	not_correct	not_limited
butted	goodwill	not_beauty	not_cry

5.2.3 Post hoc analysis and p-value adjustment

The standardized residuals from a $R \times C$ contingency table are computed as follows:

$$e_{rc} = \frac{f_{rc} - E_{rc}}{\sqrt{E_{rc}}}$$

where f_{rc} is the observed frequency and $E_{rc}$is the expected frequency in null hypothesis at position row R column C. P-values are then calculated from the distribution of square standardized residuals:

$$e_{rc}^2 \sim \chi^2(1).$$

Using Holm method, adjusted p-values are expressed as follows:

$$\tilde{p}_i = \min(\max_{j \leq i}\{(m - j + 1)p_j\}, 1)$$

5.3 Cognition difference in before and after the investigation-reveal border

5.3.1 Model Fit

A preview of the new LDA model components is shown as Table 8:

Table 8: Preview of the New Cognition Topic Dictionary

PRIMARY	SECONDARY	EMOTIONS
stranger	laboratory	knife
poison	carriage	loveday
hazell	prisoner	robbery
brown	declared	marriage
path	clock	murderer
drive	justice	wound
necklace	manager	kill
west	guard	art
passage	record	happy
revolver	copy	loved
violet	everyone	wonderful
arose	interview	frightened
crowd	inquiry	weapon
lane	moments	laugh
south	whilst	alarm
ship	plans	famous
sea	promise	greater
witness	inquiries	dreadful

PRIMARY	SECONDARY	EMOTIONS
gate	consider	relief
walls	excuse	struggle

5.4 Dictionaries

The Lexicoder Sentiment Dictionary is available at `http://www.snsoroka.com/data-lexicoder/`.

WordNet: An Electronic Lexical Database is available from MIT Press at `http://mitpress.mit.edu/books/wordnet/`.

Colin Martindale's English Regressive Imagery Dictionary is available at `https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/`.

References

Beasley, T. M. & Schumacker, R. E. (1995), 'Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures', *The Journal of Experimental Education* **64**(1), 79–93.

Fellbaum, C. (2006), 'Wordnet(s)', *Encyclopedia of Language & Linguistics* p. 665–670.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics* pp. 65–70.

Lu, B., Ott, M., Cardie, C. & Tsou, B. K. (2011), Multi-aspect sentiment analysis with topic models, *in* '2011 IEEE 11th international conference on data mining workshops', IEEE, pp. 81–88.

Martindale, C. (1975), *Romantic progression: The psychology of literary history*, Harpercollins.

Martindale, C. (1990), *The clockwork muse: The predictability of artistic change.*, Basic Books.

Porter, M. F. (2001), Snowball: A language for stemming algorithms.

Young, L. & Soroka, S. (2012), 'Affective news: The automated coding of sentiment in political texts', *Political Communication* **29**(2), 205–231.