*3L: The Southeast Asian Journal of English Language Studies – Vol 26(2): 189 – 205*
http://doi.org/10.17576/3L-2020-2602-14

# Preliminary Analysis of Malaysian Corpus of Financial English (MaCFE)

ROSLAN SADJIRIN
*Faculty of Computer and Mathematical Sciences*
*Universiti Teknologi MARA, Cawangan Pahang*

ROSLINA ABDUL AZIZ
*Akademi Pengajian Bahasa*
*Universiti Teknologi MARA, Cawangan Pahang*

NORZIE DIANA BAHARUM
*Akademi Pengajian Bahasa*
*Universiti Teknologi MARA, Cawangan Pahang*
*norziediana@uitm.edu.my*

NOLI MAISHARA NORDIN
*Akademi Pengajian Bahasa*
*Universiti Teknologi MARA, Cawangan Pahang*

MOHD ROZAIDI ISMAIL
*Akademi Pengajian Bahasa*
*Universiti Teknologi MARA, Cawangan Pahang*

ABSTRACT

*This paper presents the findings of the preliminary analysis conducted on the Malaysian Corpus of Financial English (MaCFE). MaCFE is a specialised corpus consisting of written documents compiled from banks in Malaysia and the corpus is currently housing approximately 4.3 million word tokens. The aim of the analysis was to evaluate the suitability of the texts chosen to represent the financial domain. The preliminary analysis involved generating the word list and lists of co-occurrences from MaCFE. RapidMiner Studio Educational 7.5.001 and an in-house Java programming solution was utilised to perform the analysis. The word list and lists of 50 most frequent two-word and three-word co-occurrences generated from the analysis reveal that the text compilation is representative of the financial domain in Malaysia. The study concludes by discussing the pedagogical implications of the findings.*

*Keywords: Corpus linguistics; Co-occurrences; Financial corpus; Specialised corpus; Word list*

## INTRODUCTION

The Malaysian Corpus of Financial English (henceforth MaCFE) is a specialised corpus consisting of written documents compiled from banks in Malaysia. At present, the corpus comprises 1472 electronic written documents compiled from four major banking categories in Malaysia; local Islamic bank, foreign Islamic bank, local conventional bank and foreign conventional bank. It is currently hosting approximately 4,373,230 million tokens.

This study reports the findings of the preliminary analysis on the datasets compiled in MaCFE. The aim of the analysis was to evaluate the suitability of the texts chosen to represent the financial domain. In order to determine the suitability of the text compilation, an analysis was conducted to generate the word list and lists of co-occurrence of MaCFE. In short, the study aims to fulfil the following objectives:

i. To generate a word list from MaCFE
ii. To generate 2-word and 3-word co-occurrences from MaCFE

From the findings of the analysis, the research team would be able to ascertain with a degree of confidence if the corpus has fulfilled its aim in representing the written communication of the financial domain in Malaysia. Moreover, the preliminary analysis would also help the research team understand the various techniques in corpus design and analysis, with the hope to enable them to equip MaCFE with suitable textual analysis facilities in the future.

## LITERATURE REVIEW

### SPECIALISED CORPORA

According to Sinclair (2004), a specialised corpus would contain texts that aim to represent the language of a specific discourse community and is often built to answer very specific questions (2004, as cited in Bennet, 2010). As an example "Child Language Data Exchange System" (CHILDES), which is the child component of the TalkBank system[1], was constructed to provide researchers in the study of human communication with the child language database (MacWhinney, 2000). Another example, Michigan Corpus of Academic Spoken English (MICASE), which hosts a collection of nearly 1.8 million words of transcribed speeches from the University of Michigan Ann Arbor, was designed to represent the contemporary academic speeches of the American university setting (http://www.helsinki.fi/varieng/CoRD/corpora/MICASE/background.html).

Specialised corpora can range in size and degree of specificity (Warren, 2010). They can range from as small as 53,000 words (e.g. Longman/Lancaster Spoken English Corpus) to as large as 55 million words (e.g. Longman Cambridge Corpus of Financial English). The size does not, however, determine the extent of their usefulness. Smaller corpora are easier to manage, thus, more suitable for training purposes (Yoon, 2011), they are also easier to become familiar with, interpret, construct, reconstruct plus they come with a clear pattern and clearer limits. (Aston, 1997). A small corpus is usually designed to focus on discipline-specific criteria, which caters to the needs and interests of a target discourse group (Kennedy & Miceli, 2001). As an example, the Contemporary Written Italian Corpus (CWIC) developed by Kennedy and Miceli (2001) for the teaching of writing skills in Italian, has only 570,000 words but was successfully used as a reference corpus in the integration of corpus-based writing lessons implemented by the researchers. In terms of specificity, some corpora are highly specialised, focusing on one specific text type, move, or function (Warren, 2010) such as Someya's corpus of Business Letters, consisting of only business letter samples, Hyland's Research Articles Corpus comprising only research articles (see Hyland, 1998) or The International Business Management Corpus, consisting of only research articles in the field of International Business Management sourced from two impact factor journals (Ang & Tan, 2018).

Other specialised corpora may be more varied in terms of the range of text types or language use, for instance, MICASE which contains a wide range of spoken academic texts, such as lectures, colloquia, discussions, student presentations, seminars, lab sessions, dissertation defenses, etc. or the International Corpus Network of Asian Learners of English (ICNALE) one of the largest and publicly available learner corpus consisting of controlled

[1] TalkBank is a computerised exchange system for sharing and studying conversational interactions. The system is coordinated by Brain MacWhinney from Carnegie Mellon University and can be accessed at https://childes.talkbank.org/.

speeches and essays produced by learners of English from 10 countries and areas in Asia (in Ching & Yen, 2019).

Specialised corpora are often used in English for Specific Purposes (ESP) settings (Bennett, 2010). They are most valued for their ability to provide course designers and ESP/EAP practitioners with salient lexico-grammatical features, typical choice of words (frequency), meaning nuances of near-synonyms and appropriate use of collocations unique to the domain they represent (e.g. financial, business, nursing, etc.). The series of Cambridge specialised corpora for instance (e.g. Cambridge Corpus of Legal English, Cambridge Corpus of Business English, Cambridge Corpus of Financial English) have been used for the development of business, professional and vocational teaching and learning resources (books, audio CDs, ebooks, etc.). Another notable contribution of corpora in ESP (in particular EAP) is the development of Coxhead's (2000) academic word list (AWL), which provides learners with a list of vocabulary most frequently used in academic settings. Beyond that, specialised corpora have also been developed to cater to the communication needs of practicing professionals as in the case of Hong Kong Financial Services Corpus (HKFSC) and Hong Kong Engineering Corpus (HKEC) (Warren, 2010), which are targeted for professionals in the fields of finance and engineering respectively.

Realising the pedagogical values of specialised corpora, Malaysian Corpus of Financial English (MaCFE) was built to provide ESP/EAP instructors and learners with data mined from financial institutions in Malaysia. MaCFE at the current stage consists of approximately 4.3 million words. It is comparatively smaller compared to other established financial corpora for example HKFSC, which consists of 7.3 million words (see Warren, 2010) or the Cambridge Corpus of Financial English which has over 55 million words. Nonetheless, MaCFE contains quite a wide range of text types, altogether 22 (e.g. Annual Reports, Media Releases, Fund Descriptions, Agreements, Speeches, Corporate Announcements, etc.) These text types represent the majority of most common written communication practices in the local and foreign banking institutions in Malaysia. It is hoped that the data could be expanded further, thus, benefitting those in the field of ESP/EAP and other specific fields like finance and banking.

## COMPILING MaCFE

The development of MaCFE adheres strictly to the corpus building principles posited by Sinclair (2004). According to Sinclair (2004), the contents of the corpus have to be selected according to their communicative function (external criteria) and not according to the language they may contain (internal criteria). External criteria are defined situationally and the common criteria may include text mode, text type, text domain, language variety, text locations, genre, etc., while the internal criteria are defined linguistically. Sinclair (2004) added that "corpora should be designed and constructed exclusively on external criteria". In keeping with this principle, MaCFE was designed and built according to the external criteria as summarised in Table 1. The criteria are kept small and separated from each other in line with Sinclair's fourth principle of corpus building.

TABLE 1. MaCFE external criteria

| Criteria | Description |
|---|---|
| Mode: | Written |
| Text Type: | Annual Reports, Brochures, Codes of Practice, Corporate Announcement, Circulars, Product Descriptions, Product Reports, Interim Reports, Media Releases, Ordinance, Prospectuses, etc. |
| Domain: | Financial |

| Criteria | Description |
|---|---|
| | Banks: Conventional Banks (Local & International), Islamic Banks (Local & International) |
| Language: | English |
| | Non-Native Speakers |
| Location: | Malaysia |

Due to the abstraction of natural language, it would be impossible and unrealistic for a corpus to capture all the patterns of language and to accurately represent them. In order to achieve "accurate representation" (Reppen, 2010), corpus designers can nevertheless try to sample the language that represents as much as possible the language of the discourse community they wish to represent. MaCFE has benefitted from works on existing financial corpora, in particular, Hong Kong Financial Service Corpus (HKFSC). The list of 26 text types used for HKFSC was used as a point of reference, which was then adjusted to suit the Malaysian financial system. Since Islamic and conventional banking systems operate side by side in Malaysia, the product descriptions for credit card, investment and insurance for instance, are categorised into two text types, namely Product Descriptions_Conventional and Product Descriptions_Islamic. The text types listed in Table 2 below are the most common texts that professionals in the banking sector in Malaysia would read and write, thus, the list would represent if not all, a large extent of the written language used in the sector.

TABLE 2. Text types and composition of MaCFE

| Text Type | Words | Text Type | Words |
|---|---|---|---|
| Advertisements | 45551 | Interim Reports | 248411 |
| Agreements | 10448 | Media Releases | 285764 |
| Annual Reports | 1121439 | Media Coverage | 55966 |
| Brochures | 6656 | Ordinance | 2624 |
| Bank Service Charges | 52303 | Policies | 986 |
| Corporate Announcement | 14107 | Principles | 9070 |
| Corporate Social Responsibility Reports | 18510 | Product Description_Conventional | 44946 |
| Financial Reports | 243917 | Product Description_Islamic | 326852 |
| Fund Descriptions | 95736 | Publications | 38589 |
| Fund Reports | 564524 | Speeches | 311318 |
| General Meetings | 8707 | Terms & Conditions | 866806 |
| | Total Number of Words: | 4373230 | |

The majority of the texts can be accessed via the public domain and they are mostly downloadable as entire texts, thus, meeting the sixth principle of corpus building, which states that written samples wherever possible consist of entire documents. In order to maintain homogeneity, before being downloaded, the documents were evaluated to detect obviously odd or unusual texts or texts that did not follow the standard writing convention of a particular text type. In the case of MaCFE, there was no unusual text detected. This was mainly due to the nature of the documents mined, which were formatted according to very specific guidelines and standards dictated by the financial professional bodies and the government (e.g. annual report, financial report, etc.). Once downloaded the documents went through four stages of data preprocessing: i) digitalising, (ii) data cleansing, (iii) part-of-speech tagging, and (iv) meta-linguistic annotation/markup. Digitalising involves converting documents into machine-readable texts or text file format. This was performed by either copying and pasting a downloaded document to Notepad and saving it or by opening it as a Microsoft Word document

and saving it as plain text (i.e. txt) using "save as" command. The original versions of the texts in PDF or Microsoft Word formats were stored in another repository together with the document information (e.g. year published, type of documents, http address, etc.) as reference. The digitalised data then underwent a data cleansing procedure.

MaCFE went through four mandatory data cleansing steps: (i) removing tables, (ii) removing images, (iii) correcting misspellings, and (iv) removing special characters (e.g. ^ % #). The digitalising process automatically removed any tables and images available in the documents, which then left the researchers with the tasks of correcting misspelled words and removing special characters. With the aid of Microsoft Word spell checker, the misspelled words were identified and corrected. The special characters were removed using a computerised system written in Java, developed specifically for this purpose. The next step in the preprocessing stage was part-of-speech (POS) tagging, which was done using an automated POS tagger developed by Toutanova and Manning (2000). The final step involved annotating meta-linguistic mark-up (e.g. text type, year published, gender of author, etc.) to the documents. This was performed manually using a system developed by the first author.

## TEXT PROCESSING TOOL AVAILABLE ON MaCFE

Currently, MaCFE is equipped with a built-in concordance tool that allows for word-level or phrase-level queries to be administered. A concordance is a formatted version or display of all the occurrences or tokens of a particular type in a corpus (Kennedy, 1998), or in the literary sense, concordance is an index that provides additional context for word usage (Wattenberg & Viégas, 2008). In the case of MaCFE, the concordance's index is in a form of reference (i.e. logical memory address of a computer), where the word-form is located in the program. The concordancer built for MaCFE uses the Key Word in Context (KWIC) format, where the form queried is displayed in the centre of the concordance line flanked by the context on either side. The concordancer allows for the length of the context to be determined by the user to a maximum of 12 words on either side.

Table 3 below is a sample of concordance lines generated for a randomly selected term *capital*. The *n-Words* in Table III refers to the number of tokens displayed before and after the term. The number of words displayed can be customised according to the user's needs and preferences.

TABLE 3. Word concordances

| n-Words Before | Term | n-Words After |
|---|---|---|
| … any person acting in concert with such person; capital funds, means paid-up | capital | and reserves, and includes, for the purposes of sections 37 and 61,… |
| … license. licensed business without the written consent of the Minister if its | capital | funds unimpaired by losses or otherwise are less than the minimum amount … |
| … unimpaired by losses or otherwise are less than the minimum amount of | capital | funds to be maintained by licensed institutions as may be prescribed by … |
| … losses have been incurred by the institution which Act A954. reduce its | capital | funds to an extent that the institution is no longer able to … |
| … and promotion framework. The Bank has also continued to build its human | capital | and strengthen its talent pipeline. In line with the demand for skills … |

# METHODOLOGY

## CORPUS DATA

The analysis involved a total number of 2957822 word tokens from 1065 texts. The texts were extracted from 12 types of documents, namely (1) Advertisement, (2) Corporate Announcement, (3) Annual Report, (4) Corporate Social Responsibility, (5) Financial Report, (6) General Meeting, (7) Interim, (8) Media Release, (9) Product Description, (10) Publication, (11) Speech and (12) Terms & Conditions.

## COMPUTATIONAL TOOLS

Two tools were employed to support the data analysis and they were:

1. RapidMiner Studio Educational 7.5.001: This tool was utilised to help obtain the word list and 2-word and 3-word collocations. RapidMiner Studio Educational 7.5.001 is an advanced tool for conducting data mining workflows for various tasks such as data mining, text mining, text processing and optimisation to generate word lists, word occurrences, document occurrences and n-grams (bi-gram and tri-gram).
2. An in-house Java programming solution: This program was specifically written to help the research team in the computation of word frequency and to automatically discover the association of n-gram tokens in the dataset.

# FINDINGS

## WORD LIST

The word list for MaCFE was generated using RapidMiner Studio Educational 7.5.001 and an in-house Java programming solution. The methodology employed by Verma and Gaur (2014) and Shterev (2013) in obtaining word lists was adapted in this study. The operators utilised were in the following orders:

1. Transform Cases: This operator transforms all characters into lowercase.
2. Tokenise (mode: non-letters): Splits text document containing non-letters into a single token.
3. Tokenise (mode: linguistic sentences; language: English): Splits text document containing linguistics sentences into single word token.
4. Tokenise (mode: linguistic tokens; language: English): Splits word token into a single character.
5. Tokenise (mode: specify character): Splits word token into a single character with a specified delimiter
6. Filter Special Characters: Removes special characters (e.g. !,Φ, Σ, ∅)
7. Filter Stopwords (English): Removes tokens that are English stopwords[2] (refer to Appendix C for the list of stopwords)

[2] Refer to frequent words such as determiners, conjunctions, preposition, pronouns and some verbal forms (e.g. *is*) that do not bear much meaning and they represent "noise" in the retrieval process and could damage retrieval result (Indurkhya & Damerau, 2010, p.458)

After performing the actions above, a list that contains three columns, namely Attribute Name, Total Occurrences and Document Occurrences was generated as presented in Table IV below. Due to the limitation of space Table 4 only presents 100 of the highest-ranking terms in the word list, the full list can be viewed in Appendix B.

TABLE 4. Word list generated from MaCFE

| Attribute Name | Total Occurrences | Document Occurrences | Attribute Name | Total Occurrences | Document Occurrences |
|---|---|---|---|---|---|
| bank | 38086 | 1000 | domestic | 3019 | 312 |
| financial | 20048 | 754 | sector | 2991 | 428 |
| customer | 18418 | 270 | foreign | 2978 | 396 |
| group | 14801 | 275 | tax | 2978 | 228 |
| account | 12738 | 399 | statement | 2959 | 236 |
| credit | 11134 | 481 | higher | 2932 | 377 |
| risk | 10376 | 383 | international | 2923 | 636 |
| card | 8399 | 176 | balance | 2915 | 296 |
| management | 7167 | 462 | end | 2871 | 419 |
| million | 6839 | 273 | annual | 2841 | 355 |
| growth | 6551 | 418 | system | 2834 | 400 |
| banking | 6233 | 547 | shariah | 2769 | 220 |
| cardholder | 6086 | 106 | further | 2708 | 465 |
| market | 5903 | 480 | continued | 2707 | 312 |
| business | 5597 | 545 | securities | 2697 | 260 |
| year | 5572 | 497 | basis | 2696 | 322 |
| capital | 5489 | 390 | customers | 2624 | 280 |
| services | 5241 | 503 | performance | 2606 | 299 |
| time | 5240 | 420 | strong | 2604 | 337 |
| income | 5169 | 404 | instruments | 2548 | 198 |
| cash | 4921 | 273 | liabilities | 2537 | 132 |
| conditions | 4916 | 421 | expected | 2580 | 359 |
| committee | 4785 | 198 | agrees | 2493 | 102 |
| terms | 4754 | 408 | made | 2468 | 330 |
| value | 4726 | 339 | recognised | 2439 | 137 |
| interest | 4446 | 427 | use | 2424 | 316 |
| assets | 4427 | 357 | subject | 2421 | 250 |
| information | 4303 | 432 | industry | 2418 | 364 |
| rate | 4194 | 402 | equity | 2398 | 217 |
| financing | 4189 | 401 | debt | 2395 | 474 |
| loss | 4023 | 323 | held | 2385 | 280 |
| due | 3965 | 492 | policy | 2380 | 332 |
| loans | 3880 | 299 | directors | 2380 | 119 |
| payment | 3818 | 235 | share | 2373 | 244 |
| date | 3817 | 259 | demand | 2340 | 367 |
| board | 3749 | 223 | continue | 2321 | 398 |
| global | 3672 | 431 | advance | 2318 | 174 |
| including | 3629 | 429 | products | 2311 | 359 |
| profit | 3548 | 264 | deposit | 2311 | 292 |

| Attribute Name | Total Occurrences | Document Occurrences | Attribute Name | Total Occurrences | Document Occurrences |
|---|---|---|---|---|---|
| amanah | 3570 | 164 | funds | 2302 | 341 |
| amount | 3493 | 328 | available | 2267 | 285 |
| finance | 3485 | 540 | risks | 2266 | 275 |
| investment | 3468 | 398 | changes | 2246 | 239 |
| economic | 3395 | 429 | asset | 2239 | 216 |
| net | 3325 | 306 | executive | 2010 | 113 |
| period | 3340 | 469 | December | 2208 | 246 |
| billion | 3270 | 403 | economy | 2194 | 340 |
| based | 3189 | 465 | company | 2106 | 230 |
| transactions | 3167 | 313 | applicable | 2062 | 200 |
| total | 3141 | 404 | loan | 2009 | 207 |

*Attribute Name*: Contains a set of word tokens extracted from the text collection
*Total Occurrences*: Contains the total number of occurrences of each token in a whole text collection
*Document Occurrences*: Contains the total number of documents in which the token was distributed

As shown in Table IV the word list generated from MaCFE contains terms associated with the financial/banking sector for example *bank, customer, group, financial, risk, credit, card, cardholder, million, management, growth, committee, banking, market, capital, cash,* etc. In general, the lexical content of MaCFE reveals the "aboutness" (Scott & Tribble, 2006) of the corpus or in other words paints the picture of its content. It is apparent that the list only contains terms that are associated with the financial domain.

It is also interesting to note that the list has captured two prominent Islamic banking terms (i.e. *amanah, shariah*), which are generally associated with Islamic banking products, services and reports. "Amanah" is an Arabic word defined as trustworthiness, loyalty, faithfulness, integrity, and honesty. In the context of Islamic banking in Malaysia, the term is often adopted for product naming purposes for example "CID Amanah Finance" or "Amanah Insurance", indicating their status as Islamic banking products. "Shariah" refers to "the religious law of Islam is seen as the expression of God's command for Muslims and, in an application, constitutes a system of duties that are incumbent upon all Muslims by virtue of their religious belief." (El Shamsy & Coulson, 2019). The term is most commonly connected to reporting of Shariah compliance requirements, which is a prerequisite for ensuring the legitimacy of Islamic financial products and services (Centre for Shariah Reference in Islamic Finance, 2017) as part of Shariah governance of Islamic banks and their subsidiaries.

In order to further ascertain the representativeness of MaCFE to the financial domain, its word list was also compared to that of the Hong Kong Financial Service Corpus (HKFSC). Table 5 below displays 40 of the terms extracted from the word list of HKFSC. For comparison purposes, stopwords were removed from HKFSC's list (for the full list visit HKFSC website at http://rcpce.engl.polyu.edu.hk/HKFSC/mfw.htm.).

TABLE 5. Word list of HKFSC

| Attribute Name | Total Occurrences | Percentage | Attribute Name | Total Occurrences | Percentage |
|---|---|---|---|---|---|
| company | 30467 | 0.46% | value | 11335 | 0.17% |
| financial | 23788 | 0.36% | net | 11318 | 0.17% |
| group | 23195 | 0.35% | December | 12087 | 0.16% |
| shares | 22316 | 0.34% | capital | 12063 | 0.16% |
| million | 19477 | 0.29% | property | 10650 | 0.16% |
| limited | 17098 | 0.26% | services | 10325 | 0.16% |

| Attribute Name | Total Occurrences | Percentage | Attribute Name | Total Occurrences | Percentage |
|---|---|---|---|---|---|
| share | 16016 | 0.24% | period | 10284 | 0.16% |
| business | 15881 | 0.24% | income | 10268 | 0.16% |
| year | 15810 | 0.24% | securities | 9715 | 0.15% |
| assets | 13877 | 0.21% | bank | 9442 | 0.14% |
| investment | 13664 | 0.21% | amount | 8817 | 0.13% |
| interest | 13257 | 0.20% | insurance | 8777 | 0.13% |
| total | 12791 | 0.19% | shareholders | 8645 | 0.13% |
| management | 12742 | 0.19% | development | 8525 | 0.13% |
| fund | 12610 | 0.19% | China | 8390 | 0.13% |
| exchange | 11962 | 0.18% | companies | 8383 | 0.13% |
| market | 11451 | 0.17% | information | 8160 | 0.12% |
| new | 11438 | 0.17% | offer | 8145 | 0.12% |
| directors | 11425 | 0.17% | years | 8050 | 0.12% |
| date | 11415 | 0.17% | profit | 7985 | 0.12% |

As can be seen from Table 4 and 5, both lists contain terms associated with the financial sector such as *investment*, *market*, *assets*, *value*, *securities*, etc. More importantly, Table V reveals that 28/40 (70%) of the terms occurring in HKFSC's word list are also available in MaCFE's list. The similarity in the lexical content of the corpora suggests quite clearly that they have similar text compilation (i.e. financial sector) and that MaCFE, like HKFSC, is also representative of the financial English in general. It is noted that there are differences in the total occurrence of these terms in both corpora, which indicate influence from the size of corpora and range of text selection.

## CO-OCCURRENCES OF TERMS

The preliminary analysis also involved analysing the corpus dataset for two-word and three-word co-occurrences. RapidMiner Studio Educational 7.5.001, which is also equipped with n-gram models, was employed to perform the analysis. N-gram model is a model of the probability distribution over n-letters, n-words, n-syllables or other units. The model is extensively used in text mining and natural language processing task such as text categorisation (Cavnar, Trenkle, & Mi, 1994), stemming or lemmatisation for improving retrieval accuracy (Mayfield & McNamee, 2003), text similarity measurement (Kondrak, 2005), feature selection for authorship identification (Houvardas & Stamatatos, 2006), handwriting word recognition (Poznanski & Wolf, 2016), and sentence prediction (Bickel, Haider, & Scheffer, 2005).

In this study, the n-gram model was utilised to detect and present word co-occurrences/collocation. Co-occurrence refers to the statistical tendency for words to co-occur, for example, the noun *deal* tends to co-occur with *big*, *good*, and *great* (e.g. *big deal*, *good deal*, *great deal*) (Bennett, 2010). The analysis utilised two types of n-gram models namely, bi-gram and tri-gram models available in the RapidMiner Studio Educational 7.5.001.

For the association of n-gram tokens, besides utilising the n-gram model in RapidMiner Studio Educational 7.5.001, a Java program was developed to randomly pick n-grams from the collection and associate the suffix of the n-grams to the prefix of the other n-grams randomly, for example:

$$_a(\text{N-Grams})_b <> _b(\text{N-Grams})_c <> _c(\text{N-Grams})_d$$

The first n-gram has prefix token *a* ($t_a$), and suffix token *b* ($t_b$), the second n-gram has prefix token *b* ($t_b$) and suffix token *c* ($t_c$), while the third n-gram has prefix token *c* ($t_c$) and suffix token *d* ($t_d$). Therefore, the association will be as follows:

$$t_at_b <> t_bt_c <> t_ct_d$$

In this paper, the association of n-grams for the bi-gram model was divided into three levels, which are (1) association of three n-grams, (2) association of four n-grams, and (3) association of five n-grams, while the tri-gram model was divided into two levels, which are (1) association of two n-grams, and (2) association of three n-grams. The association of n-grams in this study did not intend to summarise the text documents or the latency of information in the corpus, but rather to examine the applicability of term association for prefix and suffix of the n-grams

## CO-OCCURRENCES OF TERMS USING BI-GRAM MODEL

Table 6 below presents the list of 50 two-word occurrences in the MaCFE obtained using the bi-gram model in the RapidMiner Studio Educational 7.5.001:

TABLE 6. List of 50 two-word co-occurrences in the MaCFE

| Attribute Name | Total Occurrences | Attribute Name | Total Occurrences |
|---|---|---|---|
| credit card | 4902 | interest rate | 784 |
| fair value | 3149 | personal data | 783 |
| risk management | 2128 | domestic demand | 767 |
| financial statements | 2878 | operational risk | 772 |
| Islamic finance | 1751 | capital adequacy | 757 |
| financial instruments | 1696 | advance account | 741 |
| credit risk | 1607 | current account | 728 |
| financial institutions | 1432 | International reserves | 719 |
| foreign currency | 1375 | accounting policies | 713 |
| financial assets | 1357 | income statement | 693 |
| eligible cardholder | 1204 | asset liabilities | 691 |
| financial year | 1145 | corporate governance | 685 |
| financial system | 1017 | generic terms | 683 |
| Islamic financial | 1044 | banking system | 677 |
| financial services | 1011 | financial position | 638 |
| cash flows | 903 | interest income | 633 |
| balance sheet | 837 | saving account | 630 |
| financial reporting | 837 | financial markets | 623 |
| third party | 830 | private sector | 616 |
| debt securities | 818 | Shariah committee | 611 |
| year ended | 830 | financial stability | 607 |
| covered person | 804 | global financial | 589 |
| foreign exchange | 801 | profit rate | 589 |
| comprehensive income | 794 | financial sector | 588 |
| Islamic banking | 787 | maturity date | 574 |

The list consists primarily of lexical co-occurrences or also known as lexical collocations[3] (Benson, Benson & Ilson, 1986), which refer to co-occurrences involving only content words such as noun + noun, (e.g. *saving account*) and adjective + noun (e.g. *covered person*). Generally, most of the co-occurrences comprise terms with specific reference to the financial sector (e.g. *financial services, profit rate, credit risk*, etc.). Interestingly, several co-occurrences related to the Islamic banking system are also available in the list; they include *Islamic finance, Islamic financial, Islamic banking and Shariah committee,* which to an extent reflect the current state of banking reality in Malaysia, where the Islamic banking system is a major part of.

CO-OCCURRENCES OF TERMS USING TRI-GRAM

In generating the three-word co-occurrences the data underwent a tri-gram analysis using RapidMiner Studio Educational 7.5.001. Table 7 presents the 50 most frequent three-word co-occurrences generated from the analysis.

TABLE 7. List of 50 three-word co-occurrences in the MaCFE

| Attribute Name | Total Occurrences | Attribute Name | Total Occurrences |
|---|---|---|---|
| as well as | 571 | in respect of | 208 |
| is based on | 409 | at the end | 207 |
| will continue to | 370 | the development of | 203 |
| the end of | 302 | the international reserves | 202 |
| in line with | 299 | ensure that the | 200 |
| in accordance with | 296 | issued by the | 187 |
| to ensure that | 288 | the amount of | 193 |
| in terms of | 287 | issued by bank | 187 |
| part of the | 282 | are subject to | 185 |
| terms and conditions | 279 | performance of the | 185 |
| as a result | 278 | fees and charges | 181 |
| is expected to | 273 | the Malaysian economy | 180 |
| a result of | 256 | the purpose of | 178 |
| in addition to | 255 | continue to be | 172 |
| in accordance with | 250 | in order to | 169 |
| in the financial | 244 | as part of | 167 |
| due to the | 243 | changes in the | 163 |
| end of the | 240 | at the time | 163 |
| the right to | 234 | the implementation of | 163 |
| the use of | 233 | arising from the | 163 |
| in relation to | 228 | in which the | 160 |
| in the event | 223 | relating to the | 158 |
| finance month of | 216 | the global economy | 156 |
| products and services | 216 | is based on | 155 |
| the provision of | 209 | is sufficient to | 154 |

The three-word co-occurrences are mainly formulaic sequences or fixed strings of words with specific functions and use in the speech production (Wood, 2006) such as *as a*

[3] Benson, Benson and Ilson (1986) categorised collocations under lexical and grammatical collocations. Lexical collocations are formed with only content words (i.e. noun+noun, verb+noun, adjective+noun), while grammatical collocations are made up of a content word and a function word (e.g. verb+preposition, preposition+noun)

*result, in addition to, due to the, in the event, in order to,* etc. These sequences are also referred to as multiword expressions (MWEs), which by definition refer to "units that are decomposed of more than one (space-separated) lexical unit and display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin & Kim, 2010, p.3). The list also contains co-occurrences involving financial terms such as *terms and conditions, products and services, fees and charges, the international reserve, the global economy, issued by bank,* etc., which signify that the texts compiled for MaCFE to an extent reflect the written discourse used in the financial sector.

## CONCLUSION

Several conclusions can be derived from the preliminary analysis of MaCFE. First and foremost, the analysis of word list administered enables the research team to identify highly frequent words such as *financial, customer, account, credit, management, million, capital,* and *rate* (refer to Table IV) and co-occurrences such as *credit card, interest rate, private sector, financial year, the global economy, issued by bank* (refer to Table VI and VII), which are specifically associated with the finance industry. Based on the word list generated from this study and its comparison with the HKFSC's word list, it is apparent that the text collection for MaCFE is representative of the financial sector in Malaysia.

Second, the lists also include key Islamic finance terms and co-occurrences such as *amanah, Shariah Islamic finance, Islamic financial, Islamic banking* and *Shariah committee,* which depict the reality of banking sector in Malaysia, where Islamic banking is a major part of the financial norms. Since the enactment of the Islamic Banking Act 1983, Islamic banking has become an integral part of the banking industry in Malaysia. Hence, it is not surprising for words such as *Islamic* and *Shariah* to be captured in the lists. With the additional value rested on its collection of Islamic banking documents, MaCFE's research prospect can be expanded to include studies on Islamic banking in general and the language patterns of written communication in the Islamic financial institutions in Malaysia particularly.

Finally, the findings from the preliminary analysis have important pedagogical implications. The list of financial terms and co-occurrences from MaCFE is a valuable input in designing course materials for teaching finance-specific reading and writing in the ESP/EAP setting. First, the highly frequent words and co-occurrences can be included and emphasised in the reading component of financial English courses. Exercises like gap-filling or matching that create opportunities for learners to explore the meaning and use of words and phrases in context can be developed from the word list and lists of co-occurrences. Moreover, the inclusion of these terms and phrases in the reading texts would familiarise learners to their usage pattern in the texts read/written by financial professionals. Second, material developers can utilise the two-word and three-word co-occurrences in MaCFE in preparing learners for technical writing relevant to financial English such as reports or reviews. The co-occurrences can be explicitly taught and their usage exemplified by extracting authentic samples from the corpus. Collocation/co-occurrences are often linked with native-like lexical accuracy and fluency (Nation & Webb, 2011) and exposing learners to discipline-specific co-occurrences which can enhance learners' competency not only in technical writing, but also can increase their general fluency in the English language.

## ACKNOWLEDGMENT

This study was funded by Ministry of Higher Education (Malaysia) and Universiti Teknologi MARA (UiTM) under Research Acculturation Grant Scheme (RAGS) (RAGS/1/2014/SSI01/UITM/2).

REFERENCES

Ang, L.H. & Tan, K.H.(2018). Specificity in English for Academic Purposes (EAP): A corpus analysis of lexical bundles in academic writing. *3L: The Southeast Asian Journal of English Language Studies*, *24*(2) 82 – 94. http://doi.org/10.17576/3L-2018-2402-07

Aston, G. (1997). Small and large corpora in language learning. In B. Lewandowska-Tomaszczyk & J. P.Melia (Eds.), *Practical applications in language corpora* (pp. 51-62). Lodz, Poland: Lodz University Press.

Baldwin, T. & Su N. K. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing*, second edition. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Bennett, G. R. (2010). Using corpora in the language learning classroom: Corpus Linguistics for teachers part 1. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers Part 1*, 22. https://doi.org/10.3998/mpub.371534

Benson, M., Benson, E. & Ilson, R. (1986b). *The BBI combinatory dictionary of English: A guide to word combinations.* Amsterdam: John Benjamins.

Bickel, S., Haider, P., & Scheffer, T. (2005). Predicting sentences using N-gram language models. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 193–200. https://doi.org/10.3115/1220575.1220600

Cavnar, W. B., Trenkle, J. M., & Mi, A. A. (1994). N-gram-based text categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161–175. https://doi.org/10.1.1.53.9367

Centre for Shariah Reference in Islamic Finance, (2017) *Shariah Standard*. Retrieved Jun 16, 2020, from https://www.sacbnm.org/?page_id=3316

Ching, H. L. & Yen.L.L. (2019). Grammatical and lexical patterning of make in Asian learner writing: A corpus-based study of ICNALE: *3L: The Southeast Asian Journal of English Language Studies*. Vol 25(3): 1 – 15. *http://doi.org/10.17576/3L-2019-2503-01*

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

El Shamsy, A. & Coulson, N.J. (2019, Nov 03). *Shariah*. Encyclopædia Britannica, inc. https://www.britannica.com/topic/Shariah

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Artificial Intelligence Methodology Systems and Applications*, *4183*, 77–86. https://doi.org/10.1007/11861461_10

Hyland, K. (1998). *Hedging in scientific research articles*. John Benjamins: Armsterdam.

Indurkhya, N. & Damerau, F. J. (2010). Handbook of natural language processing (2nd. Edition). Taylor & Francis Group: Boca Raton.

Kennedy, G. (1998). *An introduction to Corpus Linguistics*. Longman, Londan and New York.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, *5*(3), 77-90.

Kondrak, G. (2005). N-gram similarity and distance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3772 LNCS, pp. 115–126). https://doi.org/10.1007/11575832_13

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk, 3rd ed.* Mahwah, NJ: Erlbaum.

Mayfield, J., & McNamee, P. (2003). Single n-gram stemming. *Proceedings of the 26th Annual International …*, *1*(240), 415–416. https://doi.org/10.1145/860435.860528

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston: Heinle.

Nation P. & Webb S. (2011). Researching and analyzing vocabulary. Boston: Heinle Cengage.

Poznanski, A., & Wolf, L. (2016). CNN-N-gram for handwriting word recognition. *Cvpr*, 2305–2314. https://doi.org/10.1109/CVPR.2016.253

Reppen, R. (2010). Building a corpus: What are the key considerations? In O'Keeffe, A and McCarthy, M. (Eds).*The Routledge handbook of Corpus Linguistics* (pp.31-37). Milton, United Kingdom:Routledge.

Roslan S., Roslina A. A., Noli M. N., Mohd Rozaidi I. & Norzie D. B. (2018). The development of Malaysian Corpus of Financial English (MaCFE). *GEMA Online® Journal of Language Studies, 18*(3), 73-100.

Scott, M. & Tribble, C. (2006). Textual patterns: Key words and corpus analysis in language education. Amsterdam: John Benjamins.

Shterev, Y. (2013). Demo: Using RapidMiner for text mining. *RapidMiner Possibility for text Mining*, *3*, 3–5.

Sinclair, J. (1991). *Corpus, concordance and collocation*. Oxford University Press: New York.

Sinclair, J. (2004). Corpus and Text — Basic principles. In *Developing linguistic corpora: A guide to good practice* (pp. 5–24).

Taylor, C. (2006). What is corpus linguistics? What the data says, 179–200.

Toutanova, K., Klein, D., & Manning, C. D. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1 (NAACL '03),* 252–259.

https://doi.org/10.3115/1073445.1073478

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* -, *13*, 63–70. https://doi.org/10.3115/1117794.1117802

Verma, T., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems*, *7*(2), 16–18.

Warren, M. (2010). Online corpora for specific purposes. *ICAME Journal*, *34*, 169–188. Retrieved from http://icame.uib.no/ij34/warren.pdf

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. In *IEEE Transactions on Visualization and Computer Graphics* (Vol. 14, pp. 1221–1228). https://doi.org/10.1109/TVCG.2008.172.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review*, *63*(1), 13-33. https://www.utpjournals.press/doi/pdf/10.3138/cmlr.63.1.13

Yoon, H. (2011). Concordancing in L2 writing class. An overview of research and issues. *Journal of English for Academic Purposes 10*, 130-139.

Zimmermann, T., & Weißgerber, P. (2004). Preprocessing CVS data for fine-grained analysis. *Proc. MSR*, 2–6. https://doi.org/10.1049/ic:20040466

## APPENDIX A: LIST OF THE 100 SAMPLE OF SELECTED WORDS TOKENS AND ITS POS TAGSET

INCLUSIVE OF ENGLISH STOPWORDS

| Word Tokens | TF | IDF | TF-IDF | Word Tokens | TF | IDF | TF-IDF |
|---|---|---|---|---|---|---|---|
| the_DT | 0.08821 | 1 | -1.05447 | was_VBD | 0.00215 | 1.28149 | -2.55987 |
| of_IN | 0.03904 | 1.00082 | -1.40819 | year_NN | 0.00201 | 1.33099 | -2.57374 |
| and_CC | 0.03677 | 1.00369 | -1.43288 | time_NN | 0.00189 | 1.4041 | -2.5772 |
| to_TO | 0.03363 | 1.00286 | -1.47209 | income_NN | 0.00186 | 1.42097 | -2.57794 |
| in_IN | 0.02447 | 1.05422 | -1.5885 | it_PRP | 0.00199 | 1.30885 | -2.58494 |
| or_CC | 0.01372 | 1.24992 | -1.76589 | business_NN | 0.00201 | 1.29095 | -2.58506 |
| a_DT | 0.01374 | 1.06545 | -1.83456 | our_PRP$ | 0.00171 | 1.5128 | -2.58645 |
| bank_NN | 0.0137 | 1.02735 | -1.85144 | value_NN | 0.0017 | 1.49715 | -2.59417 |
| for_IN | 0.01192 | 1.05747 | -1.89961 | services_NNS | 0.00189 | 1.32578 | -2.60204 |
| by_IN | 0.01006 | 1.09134 | -1.95966 | conditions_NNS | 0.00177 | 1.40307 | -2.60524 |
| customer_NN | 0.00663 | 1.59599 | -1.97565 | also_RB | 0.00211 | 1.16225 | -2.60969 |
| is_VBZ | 0.00984 | 1.03436 | -1.99235 | these_DT | 0.00185 | 1.32492 | -2.6112 |
| any_DT | 0.00708 | 1.4374 | -1.99252 | terms_NNS | 0.00171 | 1.41669 | -2.61559 |
| on_IN | 0.00913 | 1.08287 | -2.00517 | assets_NNS | 0.00159 | 1.47468 | -2.62912 |
| as_IN | 0.00895 | 1.04418 | -2.02948 | amanah_VBP | 0.00129 | 1.81251 | -2.63298 |
| be_VB | 0.00774 | 1.16881 | -2.0437 | fair_JJ | 0.00115 | 2.01875 | -2.63546 |
| group_NN | 0.00533 | 1.58802 | -2.07278 | payment_NN | 0.00137 | 1.65628 | -2.64296 |
| financial_JJ | 0.00721 | 1.14998 | -2.08117 | board_NN | 0.00135 | 1.67905 | -2.64495 |
| are_VBP | 0.00671 | 1.14142 | -2.11617 | interest_NN | 0.0016 | 1.39692 | -2.65078 |
| that_IN | 0.0064 | 1.15345 | -2.132 | date_NN | 0.00137 | 1.61405 | -2.65429 |
| with_IN | 0.00663 | 1.07943 | -2.14547 | loss_NN | 0.00145 | 1.51815 | -2.65806 |
| account_NN | 0.00458 | 1.42638 | -2.18459 | loans_NNS | 0.0014 | 1.55168 | -2.66429 |
| at_IN | 0.006 | 1.05054 | -2.20058 | information_NN | 0.00155 | 1.39187 | -2.66656 |
| from_IN | 0.005 | 1.09895 | -2.26028 | rate_NN | 0.00151 | 1.42312 | -2.66806 |
| risk_NN | 0.00373 | 1.44415 | -2.26828 | financing_VBG | 0.00151 | 1.42421 | -2.66824 |
| redit_NN | 0.00401 | 1.34521 | -2.26849 | more_RBR | 0.00167 | 1.28072 | -2.67071 |
| card_NN | 0.00302 | 1.78184 | -2.26883 | eligible_JJ | 0.00113 | 1.86899 | -2.67415 |
| shall_MD | 0.00306 | 1.7486 | -2.27182 | been_VBN | 0.00159 | 1.32234 | -2.67785 |
| not_RB | 0.00393 | 1.26995 | -2.30164 | under_IN | 0.00158 | 1.32149 | -2.68089 |
| will_VB | 0.00392 | 1.17426 | -2.33666 | statements_NNS | 0.00113 | 1.83983 | -2.68194 |
| its_PRP$ | 0.00356 | 1.24703 | -2.35259 | profit_NN | 0.00128 | 1.60575 | -2.68827 |
| this_DT | 0.00376 | 1.17426 | -2.35463 | impairment_NN | 0.00102 | 1.99797 | -2.69063 |
| cardholder_NN | 0.00219 | 2.00204 | -2.35812 | per_IN | 0.00132 | 1.53181 | -2.69536 |
| which_WDT | 0.00358 | 1.21846 | -2.36003 | we_PRP | 0.00125 | 1.5896 | -2.70306 |
| other_JJ | 0.00339 | 1.2492 | -2.37361 | their_PRP$ | 0.00154 | 1.2692 | -2.70845 |
| million_CD | 0.00246 | 1.59119 | -2.40721 | due_JJ | 0.00143 | 1.33539 | -2.72008 |
| has_VBZ | 0.00319 | 1.21914 | -2.4101 | amount_NN | 0.00126 | 1.51148 | -2.72133 |
| public_NN | 0.00281 | 1.37607 | -2.41341 | net_JJ | 0.0012 | 1.54163 | -2.73416 |
| may_MD | 0.00279 | 1.37704 | -2.41501 | global_JJ | 0.00132 | 1.39287 | -2.73511 |
| an_DT | 0.00316 | 1.20846 | -2.41835 | including_VBG | 0.00131 | 1.39489 | -2.7396 |
| such_JJ | 0.00273 | 1.33982 | -2.43675 | agrees_VBZ | 0.0009 | 2.01875 | -2.74212 |
| management_NN | 0.00258 | 1.36271 | -2.45419 | tax_NN | 0.00107 | 1.66942 | -2.74744 |
| all_DT | 0.00262 | 1.2753 | -2.47571 | service_NN | 0.00115 | 1.56048 | -2.74783 |
| growth_NN | 0.00236 | 1.40617 | -2.47958 | investment_NN | 0.00125 | 1.42747 | -2.74928 |
| have_VBP | 0.0025 | 1.27071 | -2.49808 | statement_NN | 0.00107 | 1.65444 | -2.75413 |
| committee_NN | 0.00172 | 1.73068 | -2.52583 | transactions_NNS | 0.00114 | 1.53181 | -2.75807 |
| banking_NN | 0.00224 | 1.28936 | -2.53885 | liabilities_NNS | 0.00091 | 1.90678 | -2.75931 |
| market_NN | 0.00212 | 1.34611 | -2.54377 | economic_JJ | 0.00122 | 1.39489 | -2.76855 |
| capital_NN | 0.00198 | 1.43629 | -2.54719 | shariah_NN | 0.001 | 1.68493 | -2.77502 |
| cash_NN | 0.00177 | 1.59119 | -2.55015 | billion_CD | 0.00118 | 1.42205 | -2.77647 |

## APPENDIX B: LIST OF THE 100 SAMPLE OF SELECTED WORDS TOKENS AND ITS POS TAGSET

EXCLUSIVE OF ENGLISH STOPWORDS

| Word Tokens | TF | IDF | TF-IDF | Word Tokens | TF | IDF | TF-IDF |
|---|---|---|---|---|---|---|---|
| bank_NN | 0.026053 | 1.02735 | -1.572429 | statement_NN | 0.002024 | 1.654438 | -2.475117 |
| customer_NN | 0.012599 | 1.595986 | -1.696641 | transactions_NNS | 0.002166 | 1.531805 | -2.479061 |
| group_NN | 0.010125 | 1.588017 | -1.793766 | liabilities_NNS | 0.001735 | 1.906776 | -2.480293 |
| financial_JJ | 0.013714 | 1.149978 | -1.802152 | economic_JJ | 0.002322 | 1.394892 | -2.489532 |
| account_NN | 0.008713 | 1.426377 | -1.905577 | shariah_NN | 0.001894 | 1.684927 | -2.496009 |
| risk_NN | 0.007098 | 1.444151 | -1.98927 | billion_CD | 0.002237 | 1.422045 | -2.497452 |
| credit_NN | 0.007616 | 1.345205 | -1.989473 | directors_NNS | 0.001628 | 1.951803 | -2.4979 |
| card_NN | 0.005745 | 1.781837 | -1.989817 | domestic_JJ | 0.002065 | 1.533195 | -2.499452 |
| cardholder_NN | 0.004163 | 2.002044 | -2.079107 | December204_VBN | 0.001668 | 1.890629 | -2.501094 |
| million_CD | 0.004678 | 1.591187 | -2.128199 | balance_NN | 0.001994 | 1.556058 | -2.508248 |
| management_NN | 0.004903 | 1.362708 | -2.175173 | period_NN | 0.002285 | 1.356177 | -2.50885 |
| growth_NN | 0.004481 | 1.406173 | -2.200566 | finance_NN | 0.002384 | 1.294956 | -2.510455 |
| committee_NN | 0.003273 | 1.730684 | -2.246813 | total_NN | 0.002149 | 1.420968 | -2.51526 |
| banking_NN | 0.004264 | 1.289362 | -2.259841 | instruments_NNS | 0.001743 | 1.730684 | -2.520495 |
| market_NN | 0.004038 | 1.346108 | -2.26476 | securities_NNS | 0.001845 | 1.612376 | -2.526565 |
| capital_NN | 0.003755 | 1.436285 | -2.268179 | based_VBN | 0.002181 | 1.359897 | -2.527752 |
| cash_NN | 0.003366 | 1.591187 | -2.271138 | foreign_JJ | 0.002037 | 1.429654 | -2.535757 |
| year_NN | 0.003812 | 1.330993 | -2.294726 | higher_JJR | 0.002006 | 1.451008 | -2.536079 |
| time_NN | 0.003584 | 1.4041 | -2.298183 | instructions_NNS | 0.001374 | 2.108272 | -2.538222 |
| income_NN | 0.003536 | 1.420968 | -2.298922 | annual_JJ | 0.001943 | 1.477121 | -2.542025 |
| business_NN | 0.003829 | 1.290953 | -2.306047 | sector_NN | 0.002046 | 1.395906 | -2.54424 |
| value_NN | 0.003233 | 1.49715 | -2.315154 | continued_VBD | 0.001852 | 1.533195 | -2.546827 |
| services_NNS | 0.003585 | 1.325782 | -2.323026 | customers_NNS | 0.001795 | 1.580192 | -2.547239 |
| conditions_NNS | 0.003363 | 1.403068 | -2.326222 | advance_NN | 0.001586 | 1.7868 | -2.547723 |
| terms_NNS | 0.003252 | 1.416689 | -2.336579 | cent_NN | 0.001177 | 2.4041 | -2.548179 |
| assets_NNS | 0.003028 | 1.474681 | -2.350105 | basis_NN | 0.001844 | 1.519494 | -2.552494 |
| amanah_VBP | 0.002442 | 1.812506 | -2.353965 | equity_NN | 0.00164 | 1.69089 | -2.556948 |
| fair_JJ | 0.00218 | 2.018749 | -2.356448 | performance_NN | 0.001783 | 1.551678 | -2.558136 |
| payment_NN | 0.002612 | 1.656282 | -2.363942 | system_NN | 0.001939 | 1.42529 | -2.55861 |
| board_NN | 0.002565 | 1.679045 | -2.365935 | end_NN | 0.001964 | 1.405136 | -2.559161 |
| interest_NN | 0.003041 | 1.396922 | -2.371771 | executive_NN | 0.001375 | 1.974271 | -2.56631 |
| date_NN | 0.002611 | 1.61405 | -2.375274 | subject_JJ | 0.001656 | 1.62941 | -2.568888 |
| loss_NN | 0.002752 | 1.518147 | -2.379049 | strong_JJ | 0.001781 | 1.49972 | -2.573262 |
| loans_NNS | 0.002654 | 1.551678 | -2.385279 | share_NN | 0.001623 | 1.63996 | -2.574782 |
| information_NN | 0.002943 | 1.391866 | -2.387544 | deposits_NNS | 0.001541 | 1.69897 | -2.581966 |
| rate_NN | 0.002869 | 1.423124 | -2.389042 | expected_VBN | 0.001765 | 1.472255 | -2.58531 |
| financing_NN | 0.002865 | 1.424205 | -2.38923 | asset_NN | 0.001532 | 1.692896 | -2.586228 |
| eligible_JJ | 0.002154 | 1.868987 | -2.395134 | held_VBD | 0.001631 | 1.580192 | -2.588714 |
| statements_NNS | 0.002149 | 1.839829 | -2.402929 | made_VBN | 0.001688 | 1.508836 | -2.593926 |
| profit_NN | 0.002427 | 1.605746 | -2.409252 | audit_NN | 0.001298 | 1.951803 | -2.596181 |
| impairment_NN | 0.00194 | 1.997966 | -2.411618 | changes_NNS | 0.001536 | 1.648952 | -2.596295 |
| due_JJ | 0.002712 | 1.335385 | -2.441063 | use_VBP | 0.001658 | 1.527663 | -2.596353 |
| amount_NN | 0.002389 | 1.511476 | -2.442313 | further_JJ | 0.001852 | 1.359897 | -2.598758 |
| net_JJ | 0.002274 | 1.541628 | -2.455141 | December_NNP | 0.00151 | 1.636415 | -2.60702 |
| global_JJ | 0.002512 | 1.392872 | -2.456099 | deposit_NN | 0.001581 | 1.561967 | -2.607441 |
| including_VBG | 0.002482 | 1.394892 | -2.460585 | risks_NNS | 0.00155 | 1.588017 | -2.608798 |
| agrees_VBZ | 0.001705 | 2.018749 | -2.463108 | policy_NN | 0.001628 | 1.506212 | -2.61045 |
| tax_NN | 0.002037 | 1.669415 | -2.468424 | international_JJ | 0.001999 | 1.223892 | -2.611341 |
| service_NN | 0.002177 | 1.560482 | -2.468817 | available_JJ | 0.001551 | 1.572505 | -2.612869 |
| investment_NN | 0.002372 | 1.427467 | -2.470268 | applicable_JJ | 0.001411 | 1.72632 | -2.613503 |

APPENDIX C: SAMPLES OF ENGLISH STOPWORDS

| | | | |
|---|---|---|---|
| a | can | having | it |
| about | can't | he | it's |
| above | cannot | he'd | its |
| after | could | he'll | itself |
| again | couldn't | he's | let's |
| against | did | her | me |
| all | didn't | here | more |
| am | do | here's | most |
| an | does | hers | mustn't |
| and | doesn't | herself | my |
| any | doing | him | myself |
| are | don't | himself | no |
| aren't | down | his | nor |
| as | during | how | not |
| at | each | how's | of |
| be | few | I | off |
| because | for | I'd | on |
| been | from | I'll | once |
| before | further | I'm | only |
| being | had | I've | or |
| below | hadn't | if | other |
| between | has | in | ought |
| both | hasn't | into | our |
| but | have | is | ours |
| by | haven't | isn't | ourselves |