Predicting merger decision outcomes of the European Commission:

A Natural Language Processing and Machine Learning approach

By

Ma. Adelle Gia Toledo Arbo

Master of Data Science for Public Policy

2023

Supervisor: Arianna Ornaghi, PhD

Practice Partner: Compass Lexecon EMEA, Data Science Team

Word count: 7,694

Hertie School

Berlin, Germany

Abstract

This study presents a novel application of Natural Language Processing and Machine Learning to predict merger decision outcomes of the European Commission by analyzing extracted text from official merger decisions. A Support Vector Machine linear classifier was found to be effective in predicting mergers that were either approved unconditionally or subject to conditions, achieving a high recall of 84% in identifying cases with serious anticompetitive effects and a precision of 63% to avoid imposing unnecessary conditions on cases without anticompetitive effects. The model used specific keywords related to relevant markets, demonstrating its ability to capture classification nuances. Practical implementation requires a cost-benefit analysis to strike an optimal balance between recall and precision based on the Commission's specific objectives and priorities. The study provides valuable insights into the potential of building a text-based predictive model for antitrust decisions, but emphasizes the model's limitations and importance of expert judgment in ensuring a comprehensive and reliable decision-making process.

Keywords: Natural Language Processing, Machine Learning, European Commission, Antitrust, Market Competition, Merger Review

Acknowledgements

I am deeply grateful to my family and friends both in the Philippines and Berlin for their support and encouragement throughout my master's journey abroad. Thank you to my parents, Edgar and Jinna Arbo, and my sisters, Gem and Meg Arbo, who have been my pillars of strength, and to my boyfriend, Jyrus Cimatu, for his constant support and motivation. I am also grateful to my closest friends, Vincent and Janine, for helping me navigate my life in Berlin and providing me with a sense of home away from home.

I am thankful for the guidance and feedback provided by my thesis supervisor, Arianna Ornaghi, PhD, and the opportunity to work with Compass Lexecon EMEA on practical applications of NLP and ML in the antitrust field. Special thanks to the Data Science Team, especially Enrico Alemani, for their expertise and support throughout my thesis.

Lastly, I want to express my gratitude to God for the strength and wisdom to complete this thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BoW – Bag-of-words

BERT – Bidirectional Encoder Representations from Transformers

CNN – Convolutional Neural Network

CV – Cross-validation

DL – Deep Learning

DNN – Deep Neural Network

EC – European Commission

ECtHR – European Convention of Human Rights

EU – European Union

LSTM – Long-short Term Memory

M&As – Mergers and Acquisitions

ML – Machine Learning

MLP – Multilayer Perceptron

NLP – Natural Language Processing

NACE – Statistical Classification of Economic Activities in the European Community

NAICS – North American Industrial Classification System

phase2 –Reviewed for Phase 2

SIC – Standard Industrial Classification

SVM – Support Vector Machine

tf-idf – Term Frequency-Inverse Document Frequency

wc – Approved with conditions

Executive Summary

Recent developments in data science provide new opportunities for analyzing large volumes of unstructured data, such as merger decision reports. While some studies have used Natural Language Processing (NLP) and Machine Learning (ML) techniques to forecast judicial decisions, text-based prediction of merger decision outcomes remains unexplored.

This thesis aims to explore the feasibility of applying NLP and ML to predict antitrust decisions under the EU Merger Regulation 2004 and understand the European Commission's (EC) merger review process. By analyzing the language used in merger decision reports, as proxy for the facts and conditions of a proposed merger which would otherwise only be available in the confidential merger filings, this study seeks to uncover patterns and factors that contribute to the EC's decision-making process. Given the limited review period and the numerous notifications received by the EC, building a text-based predictive model can aid merger review. The findings of this study have important implications for antitrust agencies, firms, and researchers interested in understanding the drivers of merger decisions.

Results show that a Support Vector Machine (SVM) linear classifier was the best-performing model in predicting whether a merger was approved with or without conditions and it was better at predicting cases 'approved with conditions' than cases 'approved unconditionally.' It achieved a high recall of 84% identifying cases with serious anticompetitive effects to prevent potential competition harm like higher prices, decreased quality of goods and services, and reduced innovation. Practical implementation will require a thorough cost-benefit analysis to determine the optimal trade-off between recall and precision, which ultimately depends on the competition authority's objectives and priorities.

This thesis also highlights the model's limitations and challenges in practical applications, including the need for expert judgment and human analysis. Future research can improve the study's framework by collecting more data, considering the time variable when splitting the training and test sets, and exploring different machine learning models and more sophisticated deep learning algorithms.

Overall, this study opens up exciting possibilities for leveraging data science tools in antitrust decision-making as a supplement to expert judgment and analysis.

1. Introduction

Publicly accessible information, such as legal judgments, company reports, and press releases, combined with advances in data science offer a unique opportunity to conduct innovative research using Natural Language Processing (NLP) and Machine Learning (ML). A particularly rich source of textual data is found in the extensive discussions contained within official merger decision reports published by the European Commission (EC). This data can serve as proxy for the facts, conditions, and characteristics of a proposed merger, which would otherwise only be available in confidential merger filings submitted by the merging parties to the EC.

The EU Merger Regulation 2004[1] requires merging firms to notify the EC when a proposed merger with an EU dimension meets specific turnover thresholds. For example, if the "combined worldwide turnover of all the merging firms is over €5 billion" (EC, 2013, pg.1). To comply with this regulation, the merging parties submit extensive documentation to the EC, as outlined in the Notification Form.[2] These merger filings contain detailed information about the parties' business activities, market definitions, measures of concentration, and other relevant facts.

Following notification, the EC's merger review starts in a Phase I review, which takes 25 working days to assess if sufficient competitive constraints exist within the merger's relevant markets (Figure 1). It concludes with one of the two main decisions: merger is cleared, either unconditionally or subject to conditions, or recommended for Phase II review. If the merger proceeds to Phase II, the review takes an additional 90 working days[3] to conduct a more in-depth analysis, with a final decision of either approved unconditionally or conditionally, or prohibited.[4]

Merger decisions are classified based on the relevant Article of the Regulation that was applied during the review process (Table 1). For the purposes of this study, the decisions were classified into two binary labels: "reviewed for Phase 2" (*phase2*) and "approved with conditions" (*wc*), which represent the outcomes of EC merger decisions (Table 2).

[1] Council Regulation (EC) No 139/2004

[2] The Form has seven sections that require information from the merging parties.

[3] Phase I review can be extended by 10 days if the merging firms offer remedies. Phase II review can be extended by 15 to 20 working days upon request of the merging parties.

[4] The EU also has a referral mechanism in which the EC and EU Member States' competition authority can transfer the merger review between themselves in special cases.

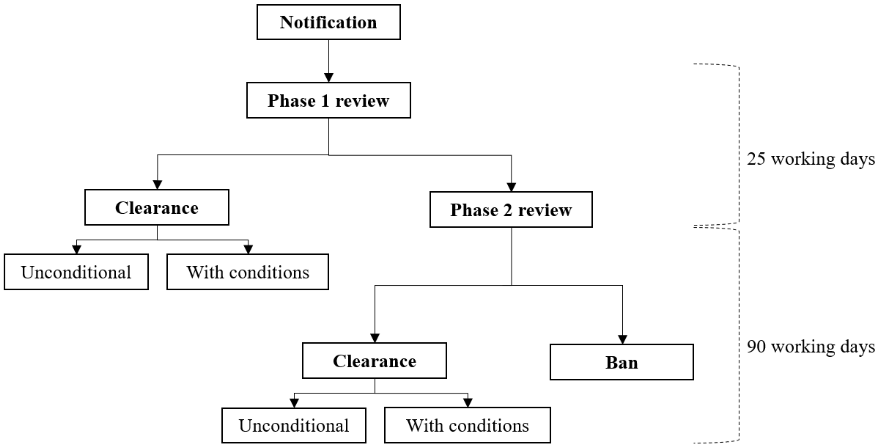


Figure 1. EU merger review process

Table 1. EC merger decision outcomes by article

Review	Article	Description
Phase 1	Art 6(1)(b)	Approved unconditionally
	Art 6(2)	Approved with conditions
Phase 2	Art 8(1)	Approved unconditionally
	Art 8(2)	Approved with conditions
	Art 8(3)	Prohibition
Referral	Art 4(4), 9(3), 22(3)	Referral mechanism

Table 2. Labels for text classification

Label	Description	Article
Reviewed for phase 2		
0	Phase 1 review	Article 6.1 and 6.2
1	Phase 2 review	Article 8.1, 8.2, 8.3
Approved with conditions		
0	Approved unconditionally	Article 6.1 and 8.1
1	Approved with conditions	Article 6.2 and 8.2

Between 2016 and 2021, an average of 13,000 merger filings were submitted annually across 58 jurisdictions globally, including the EC and competition authorities from 25 EU member states (White & Case, 2022). Since the first EU Merger Regulation in 1989,[5] the EC (2013) has received about 300 to 700 merger filings annually. In 2021, the EC made 396 merger decisions in various economic sectors.

[5] Council Regulation (EEC) No 4064/89

With numerous merger notifications and a limited review period, a text-based approach utilizing NLP and ML can be valuable in aiding merger reviews. This approach can benefit not only antitrust agencies but also firms by flagging potentially anticompetitive mergers.

Despite numerous studies on mergers and acquisitions (M&As) in finance literature, little has been done using data science in competition economics. While some researchers have used NLP and ML to analyze merger activity (Moriarty et al., 2019 and Routledge et al., 2017) and closeness of competition (Hoberg & Phillips, 2016), and others have applied these methods to predict judicial decisions like judgments of the European Court of Human Rights (ECtHR) (Aletras et al., 2016 and Medvedeva et al., 2020), predicting M&A decision outcomes remains unexplored.

This thesis aims to fill this gap by exploring the feasibility of applying NLP and ML techniques to predict merger decisions under the EU Merger Regulation 2004 and understand EC's merger review and decision-making process.

Specifically, this study aims to answer the following research questions:

1. Can an NLP and ML-based approach predict EC merger decision outcomes using extracted text from official merger decisions?
2. Which classification model is the "best" model in predicting EC merger decision outcomes?
3. What are the most important textual features in predicting EC merger decision outcomes?

The discussion of the next chapters is outlined as follows: Chapter 2 provides a review of related literature on antitrust and competition, and how data science tools have been used to analyze them; Chapter 3 focuses on the data collection, pre-processing, and balancing steps; Chapter 4 presents the methodology and text classification pipeline with NLP and ML; Chapter 5 details the experiments conducted to answer the research questions; Chapter 6 discusses the implications of results for antitrust and competition policy; Chapter 7 summarizes the key findings and conclusions, discusses the study's challenges and limitations, and suggests areas for future work.

2. Review of Related Literature

Routledge et al. (2017) were the pioneers in using text data to predict merger activity. They classified US firms as merger targets and acquirers by analyzing annual company reports and Form 10-K SEC filings' section on Management Discussion and Analysis. A similar study by Moriarty et al. (2019) utilized a larger dataset that also included the 10-K Business Description section, while (Jiang, 2021) used the entire 10-K document. These studies used ML models trained on *n-gram* representations, such as bag-of-words (BoW) and term frequency-inverse document frequency (*tf-idf*), as textual features to predict M&A activity with regularized logistic regressions.

Hoberg & Phillips (2016) also utilized 10-K product descriptions to analyze the closeness of competition of firms based on their product offerings. Cosine similarity scores were calculated using word vectors weighted by *tf-idf*, which were clustered to map firms into new industries called "text-based network industry classifications." Additionally, Morzenti (2022) used NLP techniques to analyze the relationship between non-notified horizontal mergers[6] and patenting activity in the US. Instead of *tf-idf* weighting, Doc2Vec word embeddings were used to train the corpus of US patents and calculate cosine similarity scores of patent information of the merging parties to classify horizontal mergers. Both studies showed that their methods performed better than traditional industry classifications like Standard Industrial Classification (SIC) and North American Industrial Classification System (NAICS) codes.[7]

A few studies have applied NLP and ML to predict judicial decisions. Aletras et al. (2016) formulated a binary classification task to classify whether there is a violation of Articles 3, 6, and 8 of the ECtHR using extracted text from relevant sections of published judgments. To extract textual features, an *n-gram* BoW approach was employed, and *n-grams* were clustered to create topics. Each set of textual features were used to train Support Vector Machine (SVM) classifiers, which yielded an average accuracy of 79%.

Medvedeva et al. (2020) conducted the same study but used *tf-idf* scores of *n-grams* to train SVM linear classifiers to predict whether there is a violation of 9 articles of the ECtHR, which achieved an average accuracy of 75%. They also split the training

[6] A horizontal merger is merger where "firms concerned are actual or potential competitors in the same relevant market" (EC, 2004).

[7] SIC and NAIC codes "identify a firm's primary business activity" (Washington State Department of Revenue, n.d.).

and test sets based on the case year to analyze if cases from the past can predict future cases. Their models yielded an accuracy between 58% to 68% and suggested that it is harder to predict future judgments as the gap between the training and test sets increases. Lastly, they trained a model that achieved an average accuracy of 65% using textual features derived from the surnames of judges in the Chamber, representing each judge's presence on the bench as a feature.

Virtucio et al. (2018) adapted the framework introduced by Aletras et al. (2016) to predict Supreme Court decisions for criminal cases in the Philippines. They employed the BoW approach and spectral clustering to generate textual features, which were fed as input to SVM linear classifiers and random forest classifiers. Their results showed an accuracy of 59% using a random forest classifier.

Researchers are also investigating the use of more advanced algorithms, such as Deep Learning (DL), to build predictive models using text data. Kastrati et al. (2019) proposed a document classification approach involving semantic enrichment for document representation by "using background knowledge provided by ontologies" and extracting linguistic information (e.g., synonyms) to capture the "whole conceptualization of documents." They used *tf-idf*, word embeddings, and topic modeling to generate text representations as inputs to various DL classifiers. They trained deep neural networks (DNN) like multilayer perceptron (MLP), long-short term memory (LSTM), and convolutional neural network (CNN) using the INFUSE[8] dataset and found that DNNs[9] performed better than ML models like SVM and Decision Tree.

This study aims to contribute to the presently thin literature on building text-based predictive models for legal decisions, particularly for merger review. As a starting point, it builds upon the framework proposed by Medvedeva et al. (2020) in a relatively unexplored area of research.

[8] According to ChatGPT (31 January 2023) response, INFUSE (INcorporating Feedback in Unsupervised SEmantics) is "a large-scale, unsupervised dataset designed to evaluate the performance of machine learning models in generating high-quality word representations, or word embeddings, that capture the semantic and syntactic relationships between words in a given language."

[9] MLP is a feed-forward network with all neurons in one layer fully connected to the neurons in the adjacent layer. LSTM is a recurrent neural network which has both forward and backward connections. CNN has a convolution in each layer (Kastrati et al., 2019, pg.7).

3. Data

3.1. Collecting the data

3.1.1. Data source

This study utilized cases reviewed under the EU Merger Control Regulation 2004, with notification dates between 1 May 2004 and 31 December 2022. Although merger decisions are publicly available through the EC competition search tool,[10] manually downloading each decision in PDF format over almost a 20-year period would be an arduous task. To streamline this process, a Python program that scrapes the search tool's website was created to automatically download the documents.[11]

The program was restricted to download English documents and retrieved a total of 4,855 unique merger cases as of 24 January 2023.[12] Other information obtained from the webpage were the case number, title, notification date, and NACE or economic activity concerned (Figure 2).

The initial distribution by article in Table 3 shows that majority of cases (3,194) were under approved unconditionally under Simplified Procedure, with a standardized decision of only three pages long. Such cases were excluded in this study as they only contain short descriptions of the merging parties’ business activities (Figure 3). Additionally, 59 referral cases were disregarded since they duplicated cases from other articles and did not provide final decisions. While this step diminished the dataset to 1,602 cases for parsing and pre-processing, this made the case structure consistent by removing the noise attributed to how a merger decision was written.

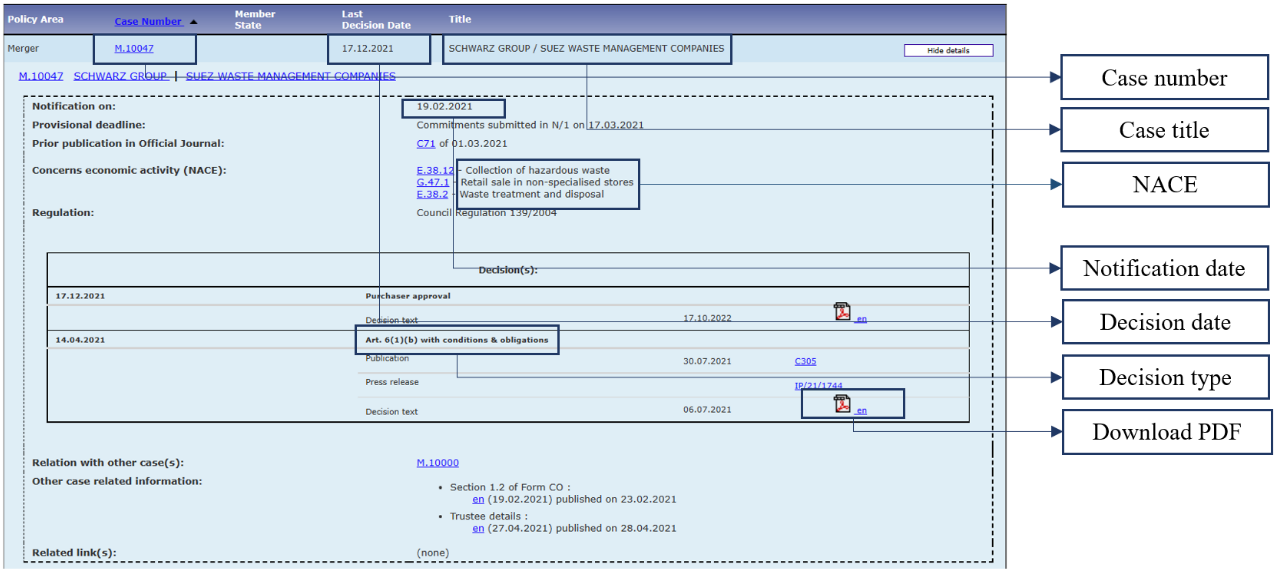
[10] https://ec.europa.eu/competition/elojade/isef/index.cfm

[11] The Github repository containing the code, data, and Python environment used in this study is available at https://github.com/adellegia/predicting-merger-decision-outcomes. Code was self-written, debugged with StackOverflow and ChatGPT, unless noted otherwise.

[12] The program tried to retrieve not only a sample but all the merger decisions under the relevant articles with an English PDF available.

Table 3. Initial distribution of merger decisions (in English) obtained from the EC Competition Search Tool on 24 January 2023

Review	Article	Description	No. of cases	Retained
Phase 1	Art 6(1)(b)	Simplified Procedure	3,194	✗
	Art 6(1)(b)	Approved unconditionally	1,285	✓
	Art 6(2)	Approved with conditions	212	✓
Phase 2	Art 8(1)	Approved unconditionally	32	✓
	Art 8(2)	Approved with conditions	63	✓
	Art 8(3)	Prohibition	10	✓
Referral	Art 4(4), 9(3), 22(3)	Referral mechanism	59	✗
Total			**4,855**	**1,602**



Policy Area	Case Number ▲	Member State	Last Decision Date	Title	
Merger	M.10047		17.12.2021	SCHWARZ GROUP / SUEZ WASTE MANAGEMENT COMPANIES	Hide details

M.10047 SCHWARZ GROUP | SUEZ WASTE MANAGEMENT COMPANIES

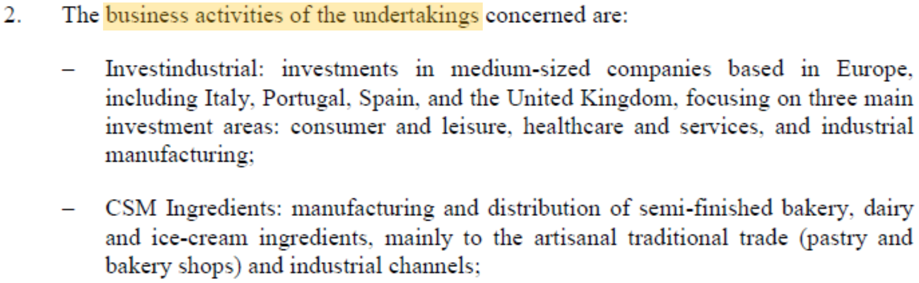
Notification on:	19.02.2021
Provisional deadline:	Commitments submitted in N/1 on 17.03.2021
Prior publication in Official Journal:	C71 of 01.03.2021
Concerns economic activity (NACE):	E.38.12 - Collection of hazardous waste G.47.1 - Retail sale in non-specialised stores E.38.2 - Waste treatment and disposal
Regulation:	Council Regulation 139/2004

Decision(s):

17.12.2021	**Purchaser approval**		
	Decision text	17.10.2022	en
14.04.2021	**Art. 6(1)(b) with conditions & obligations**		
	Publication	30.07.2021	C305
	Press release		IP/21/1744
	Decision text	06.07.2021	en

Relation with other case(s):	M.10000
Other case related information:	• Section 1.2 of Form CO : en (19.02.2021) published on 23.02.2021 • Trustee details : en (27.04.2021) published on 28.04.2021
Related link(s):	(none)

Figure 2. Scraping the EC competition search tool



2. The business activities of the undertakings concerned are:

– Investindustrial: investments in medium-sized companies based in Europe, including Italy, Portugal, Spain, and the United Kingdom, focusing on three main investment areas: consumer and leisure, healthcare and services, and industrial manufacturing;

– CSM Ingredients: manufacturing and distribution of semi-finished bakery, dairy and ice-cream ingredients, mainly to the artisanal traditional trade (pastry and bakery shops) and industrial channels;

Figure 3. Business description from a case under Article 6(1)(b) Simplified Procedure[13]

[13] Case M.10010 - INVESTINDUSTRIAL GROUP / CSM INGREDIENTS

3.1.2. Merger decision structure

A merger decision of the EC is typically written in this structure with the following sections:

- *Summary or Introduction*, consisting of the merger notification details and a brief summary with the final decision;
- *Parties and Operation*, containing the description of the merging parties and their business operations;
- *Concentration and Dimension*, containing details about the proposed transaction, including its size, structure and control, and market presence;
- *Market Definition*, identifying the relevant product and geographic markets including broad and narrow definitions for assessing the merger's competitive effects;
- *Competitive Assessment*, consisting the analysis of the impact on competition of the merger given the competitive constraints within the relevant product and geographic markets. This section primarily includes the discussion of:
    - *Theories of harm* such as horizontal effects (i.e., substantial lessening of competition), vertical effects (i.e., input foreclosure or customer foreclosure), conglomerate effects, and coordinated effects;
    - *Market shares* and quantitative analysis like diversion ratios, upward pricing pressure, and critical loss analysis;
    - *Efficiencies* that may benefit consumers post-merger;
    - *Barriers to entry* and potential new entrants and whether they are likely to enter the market in a timely and sufficient manner;
- *Remedies or Commitments*, enumerating the potential remedies imposed to the merging parties like divestitures, licensing agreements, or behavioral commitments. (Only available for cases under Article 6(2) and 8(2));
- *Conclusion*, concluding with the final decision of whether the proposed merger is approved conditionally or unconditionally, or prohibited.

The length of merger decisions and details discussed in specific sections may vary depending on the nature of the merger and the relevant market conditions. The Competitive Assessment section may include more rigorous analysis like local overlap analysis for cases with multiple local relevant geographic markets or catchment areas. Some decisions may also discuss potential impacts on innovation, consumer welfare, and public interests like data protection and privacy.

3.1.3. Text parsing and pre-processing

To automate text parsing, a program was built using Python's *pdfplumber* to extract text from 1,602 PDFs on a section-by-section basis. The program identifies subheadings by locating lines containing bold and capitalized letters and extracts the corresponding section text along with tables and footnotes (Figure 4).

Some decisions were also re-labeled based on the article number of the document's first page (Figure 5) since a few decisions were found to be mislabeled when searched and downloaded through the search tool. The language of the parsed texts was also identified to retain English texts.

Four sections were selected for the classification model, namely *Parties and Operation*, *Concentration and Dimension*, *Market Definition*, and *Competitive Assessment*. Other sections were excluded to avoid biased predictions. For instance, the *Conclusion* section in Figure 4 was excluded to prevent the model from being influenced by the decision ('not significantly impede effective competition').

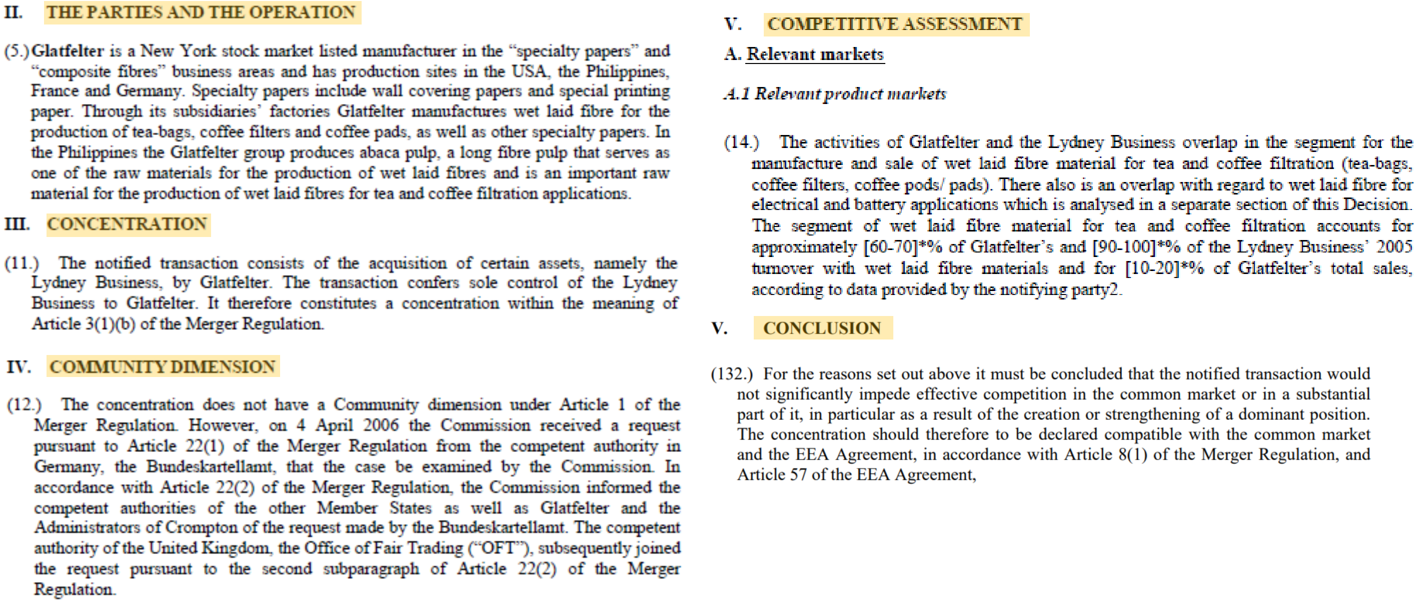
The final dataset consists of 1,583 unique merger cases with a total of 5,362 rows, corresponding to the four retained sections (Table 4).

This parsing technique was not entirely flawless, some merger decisions lacked certain sections, particularly *Market Definition* (Appendix B Table B.1). This is due to inconsistency in naming subheadings and writing merger decisions. Despite limitations, this remains the most effective method available.

Table 4. Final number of merger decisions

Review	**Article**	**Description**	**No. of cases**	**No. of rows**
Phase 1	Art 6(1)(b)	Approved unconditionally	1,281	4,398
	Art 6(2)	Approved with conditions	204	672
Phase 2	Art 8(1)	Approved unconditionally	32	94
	Art 8(2)	Approved with conditions	57	170
	Art 8(3)	Prohibition	9	28
Total			**1,583**	**5,362**

After parsing, the text data was cleaned and pre-processed by lowercasing all words, removing tags, punctuations, numeric and nonalphanumeric characters, lemmatizing words to their base form instead of stemming, and removing frequent and obvious words like the footnote in Figure 6. These steps ensured that the text data was cleaned and processed efficiently, allowing for more accurate and meaningful analysis.



II. THE PARTIES AND THE OPERATION

(5.) **Glatfelter** is a New York stock market listed manufacturer in the "specialty papers" and "composite fibres" business areas and has production sites in the USA, the Philippines, France and Germany. Specialty papers include wall covering papers and special printing paper. Through its subsidiaries' factories Glatfelter manufactures wet laid fibre for the production of tea-bags, coffee filters and coffee pads, as well as other specialty papers. In the Philippines the Glatfelter group produces abaca pulp, a long fibre pulp that serves as one of the raw materials for the production of wet laid fibres and is an important raw material for the production of wet laid fibres for tea and coffee filtration applications.

III. CONCENTRATION

(11.) The notified transaction consists of the acquisition of certain assets, namely the Lydney Business, by Glatfelter. The transaction confers sole control of the Lydney Business to Glatfelter. It therefore constitutes a concentration within the meaning of Article 3(1)(b) of the Merger Regulation.

IV. COMMUNITY DIMENSION

(12.) The concentration does not have a Community dimension under Article 1 of the Merger Regulation. However, on 4 April 2006 the Commission received a request pursuant to Article 22(1) of the Merger Regulation from the competent authority in Germany, the Bundeskartellamt, that the case be examined by the Commission. In accordance with Article 22(2) of the Merger Regulation, the Commission informed the competent authorities of the other Member States as well as Glatfelter and the Administrators of Crompton of the request made by the Bundeskartellamt. The competent authority of the United Kingdom, the Office of Fair Trading ("OFT"), subsequently joined the request pursuant to the second subparagraph of Article 22(2) of the Merger Regulation.

V. COMPETITIVE ASSESSMENT

A. Relevant markets

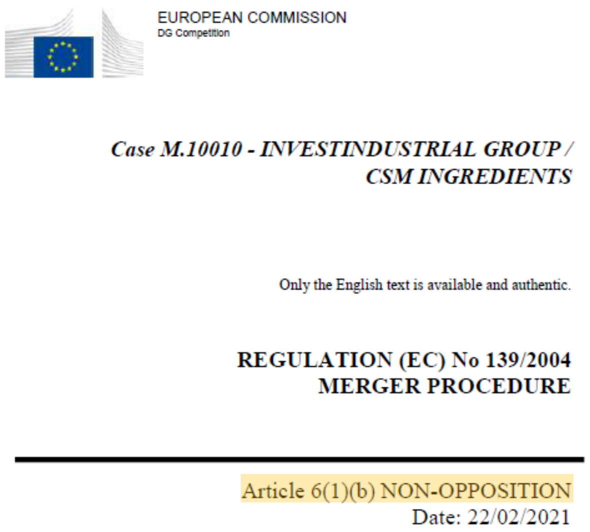
A.1 Relevant product markets

(14.) The activities of Glatfelter and the Lydney Business overlap in the segment for the manufacture and sale of wet laid fibre material for tea and coffee filtration (tea-bags, coffee filters, coffee pods/ pads). There also is an overlap with regard to wet laid fibre for electrical and battery applications which is analysed in a separate section of this Decision. The segment of wet laid fibre material for tea and coffee filtration accounts for approximately [60-70]*% of Glatfelter's and [90-100]*% of the Lydney Business' 2005 turnover with wet laid fibre materials and for [10-20]*% of Glatfelter's total sales, according to data provided by the notifying party2.

V. CONCLUSION

(132.) For the reasons set out above it must be concluded that the notified transaction would not significantly impede effective competition in the common market or in a substantial part of it, in particular as a result of the creation or strengthening of a dominant position. The concentration should therefore to be declared compatible with the common market and the EEA Agreement, in accordance with Article 8(1) of the Merger Regulation, and Article 57 of the EEA Agreement,

Figure 4. Parsing of texts on a section-by-section basis[14]



EUROPEAN COMMISSION
DG Competition

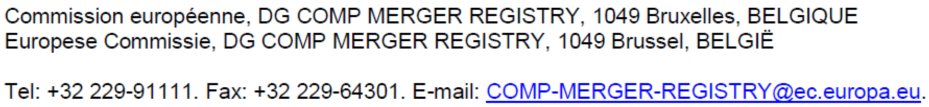
Case M.10010 - INVESTINDUSTRIAL GROUP / CSM INGREDIENTS

Only the English text is available and authentic.

REGULATION (EC) No 139/2004
MERGER PROCEDURE

Article 6(1)(b) NON-OPPOSITION
Date: 22/02/2021

Figure 5. Re-labeling documents based on first page



Commission européenne, DG COMP MERGER REGISTRY, 1049 Bruxelles, BELGIQUE
Europese Commissie, DG COMP MERGER REGISTRY, 1049 Brussel, BELGIË

Tel: +32 229-91111. Fax: +32 229-64301. E-mail: COMP-MERGER-REGISTRY@ec.europa.eu.

Figure 6. Footnote of contact details of the EC

[14] Case M.4215 – Glatfelter/ Crompton Assets

3.2. Balanced data

As most mergers were approved during Phase 1 review, the data collected in this study was highly imbalanced (Table 5). The final dataset had an imbalance factor of 15, wherein for every case that proceeded to Phase 2, 15 cases were cleared in Phase 1. Meanwhile, the imbalance ratio of cases approved unconditionally to approved with conditions was 5:1.

Balancing the class distribution is crucial to avoid biased predictions and prevent the model from solely learning the majority class. To address class imbalance, the steps below were adapted from Medvedeva et al. (2020) with *phase2* as an example:

1. Randomly remove *phase2*=0 cases to balance the distribution of labels,[15] resulting in 98 *phase2*=0 cases and 98 *phase2*=1 cases, with 1,386 *phase2*=0 cases excluded.
2. Randomly stratify based on *phase2* and article to split data into 80% training set and 20% test set,[16] ensuring that each set is representative of both *phase2*=0 and *phase2*=1 cases. For instance, the balanced training set has 79 *phase2*=0 cases and 78 *phase2*=1 cases, while the balanced test set has 20 *phase2*=0 cases and 20 *phase2*=1 cases.
3. Lastly, add the excluded *phase2*=0 cases to create an imbalanced test set with 1,406 *phase2*=0 cases and 20 *phase2*=1 cases.

Table 6 shows the class distribution after balancing and splitting the data for training and evaluation.

[15] By setting the *random seed* = 42

[16] By using *train_test_split* from *scikit-learn* and setting *random_state* = 42

Table 5. Distribution of merger decisions by *phase2* and *wc*

Label	**Description**	**No. of cases**	**No. of rows**
Reviewed for phase 2			
0	Phase 1 review	1,485	5,070
1	Phase 2 review	98	292
Total		**1,583**	**5,362**
Approved with conditions			
0	Approved unconditionally	1,313	4,492
1	Approved with conditions	261	842
Total		**1,574**	**5,334**

Table 6. Training and test data split by *phase2* and *wc*

	Balanced		**Imbalanced**
	Training set	**Test set**	**Test set**
Reviewed for phase 2			
phase2=0	79	20	1,406
phase2=1	78	20	20
Total	**157**	**40**	**1,426**
Approved with conditions			
wc=0	209	52	1,104
wc=1	208	54	54
Total	**417**	**106**	**1,158**

4. Methodology

This chapter presents the text classification pipeline that was followed in this study (Figure 7). After data preparation and balancing, the pre-processed texts were transformed into numerical form. Following the framework of Medvedeva et al. (2020), *tf-idf* scores of *n-grams* were computed to create a feature matrix using Python's *scikit-learn*. This matrix served as input to train and evaluate the machine learning models. Appendix A describes the feature matrix generation.

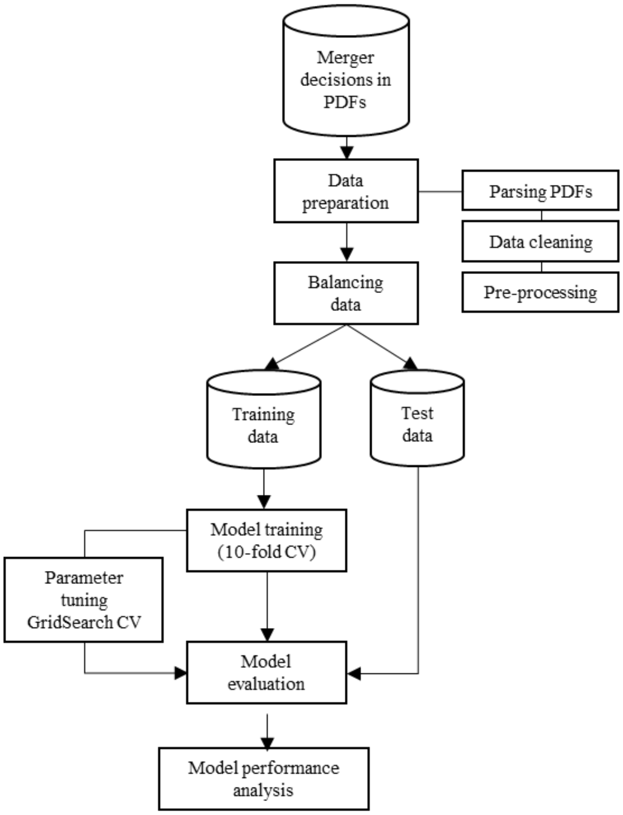


Figure 7. Machine learning for text classification pipeline

Linear Support Vector Machine was the chosen model for the classification task due to its simplicity and robustness to noisy, imbalanced, and small data (Aletras et al., 2016 and Medvedeva et al., 2020). Figure 8 shows a hypothetical example where the blue circles (negative class) and orange squares (positive class) depict the two classes of *phase2*. The goal of the linear SVM is to find the decision boundary line or hyperplane (represented by the black line) that best separates these two classes such that it maximizes the margin, which is the distance between the hyperplane and the closest data points of each class called support vectors. The algorithm works by minimizing a cost function that penalizes misclassifications while also maximizing the margin.

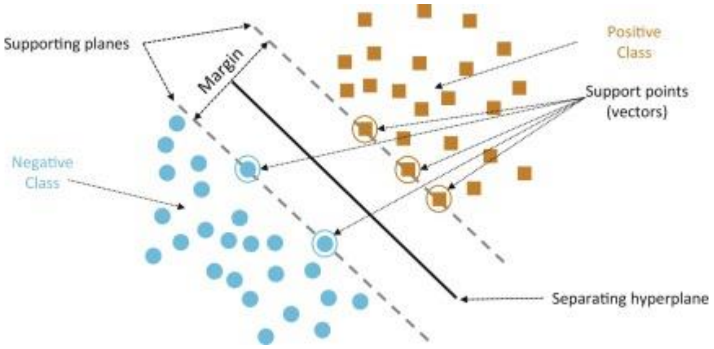


Figure 8. Illustration of SVM linear classifier *(Image from Ragab et al., 2021)*

For parameter tuning, the models were trained using a grid search method with cross-validation to identify the best parameters. The grid search was set to optimize the f1-score, which provides balance between recall and precision.

A fivefold cross-validation was conducted using *GridSearchCV* from *scikit-learn* to evaluate the model by partitioning the training data into five subsets. Each subset (1/5th) was used once as the validation set while the remaining subsets (4/5ths) were used for training. By repeating this process five times with different samples of data for training and validation, the model performance was accurately assessed and the risk of overfitting was reduced.

Once the optimal parameters were identified, the SVM linear classifier was trained with a tenfold cross-validation using the best set of parameters. The final model was used to make predictions on the held-out test set. [17]

Four evaluation metrics were used to assess model performance: recall, precision, false positive rate (FPR), and f1-score. A high recall was desired to correctly identify all relevant positive cases, but it can lower precision or the proportion of positive predictions that were actually correct, and increase FPR. Thus, balance between recall and precision was considered using their harmonic mean or f1-score.

Finally, the weights assigned to the n-grams were also examined to evaluate their importance in distinguishing the two classes. This analysis identified specific keywords and phrases that are indicative of potentially anticompetitive merger cases which require further scrutiny during Phase 2 or approval with conditions.

[17] The code for the parameter tuning, model training with tenfold cross-validation, and model evaluation was adapted from Medvedeva et al. (2020) available at https://github.com/masha-medvedeva/ECtHR_crystal_ball.

5. Experiments

Chapter 5 presents the two experiments conducted in this study, detailing the set-up, results, and discussion. Experiment 1 used words and phrases in the full text from the four merger decision sections to predict the two binary labels: *phase2* and *wc*. Experiment 2 trained models using text from each individual section.

5.1. Experiment 1: Text classification of *phase2* and *wc* using full text

5.1.1. Set-up

In Experiment 1, the linear SVM was limited to 5000 features. Parameter tuning tested all possible parameter combinations five times using random data splits for training and testing. Table 7 shows the evaluated parameters. The configuration that yielded the highest average performance was selected through cross-validation to ensure that the model generally performed well and the selection of best parameters was not influenced by chance.

GridSearchCV evaluated 864 possibilities for parameter tuning and the fivefold cross-validation trained 4,320 models for a single label. A total of 8,640 models were trained for *phase2* and *wc*.

Table 7. List of evaluated parameter values

Parameter	Values	Description
ngram_range	(1,1), (1,2), (1,3), (2,2), (2,3), (3,3)	Length of n-grams; e.g., (1,3) contains unigrams, bigrams, and trigrams
max_df	0.01, 0.025, 0.05	Exclude terms that occur as more than n% of corpus; e.g., a lower threshold will exclude most frequently occurring terms in documents
user_idf	False, True	Use inverse document frequency weighting
binary	False, True	Set term frequency to binary (all non-zero terms are set to 1)
norm	'l1', 'l2'	Norm used to normalize term vectors
stop_words	None, 'English'	Remove English stop words
C	0.1, 1, 5	Penalty term for the model

5.1.2. Results

Table 8 provides the parameter configuration selected by *GridSearchCV* for *phase2* and *wc*. For *phase2*, a combination of unigrams and bigrams achieved the best results while for *wc*, bigrams were found to be better. For both labels, the parameter search identified that removing English stop words improved model performance. These combinations of parameters were used to train the SVM linear classifiers with a tenfold cross-validation to ensure general performance and reduce sensitivity to specific cases. Compared to fivefold, tenfold cross-validation allowed for more data to be used in each fold, which is crucial due to the small data size.

Table 8. Best parameters obtained from grid search with fivefold cross-validation

Parameter	**Value**	
	phase2	***wc***
ngram_range	(1,2)	2
max_df	0.05	0.025
user_idf	FALSE	FALSE
binary	TRUE	TRUE
norm	l2	l1
stop_words	English	English
C	0.1	1

The SVM linear classifiers for predicting *phase2* and *wc* are labeled as *svm_p2* and *svm_wc*, respectively. The model *svm_p2* (Figure 9 left) achieved a high recall (above 80%) and precision (above 90%) with a low FPR (below 10%.) for both cross-validation and balanced test set. This indicates that the model accurately predicted both classes while avoiding false positives and false negatives (Figure 9 right). However, although *svm_p2* seemed to perform well, overfitting was a concern due to the small data size on which it was trained (n_{p2}=157). The robustness checks in Appendix C showed that the model performance was not consistent when different random seeds and random states (*k*=28 and *k*=70) were used in the data balancing and train-test split steps (i.e., FPR is 7-13 times larger in robustness checks).

Meanwhile, the model *svm*_wc achieved a recall of 84% and a lower precision of 63% under cross-validation, and a recall of 91% and a precision of 67% for the balanced test set (Figure 10 left). The high recall indicates that the model correctly identified the majority of positive instances, while the lower precision suggests that the model incorrectly identified some negative instances as positive, hence a high false positive rate (FPR) of about 50%.

The model *svm_wc* also predicted the classes differently with a high recall and lower precision for cases ‘approved with conditions’ (*wc*=1), and a low recall and higher precision for cases ‘approved unconditionally’ (*wc*=0) (Figure 10 right). This precision-recall tradeoff was reflected for both balanced training and test sets. More importantly, *svm_wc* was found to be robust with a consistently high performance (Appendix C).

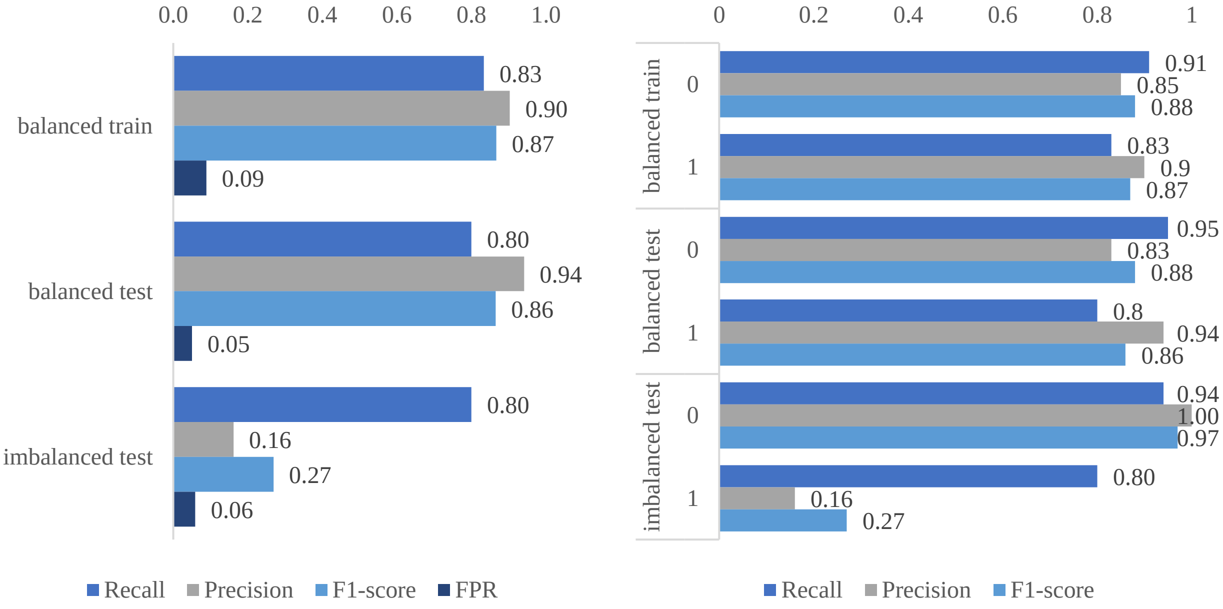


Figure 9. Performance of *svm_p2* (left) and per class (right) tenfold cross-validation and test results

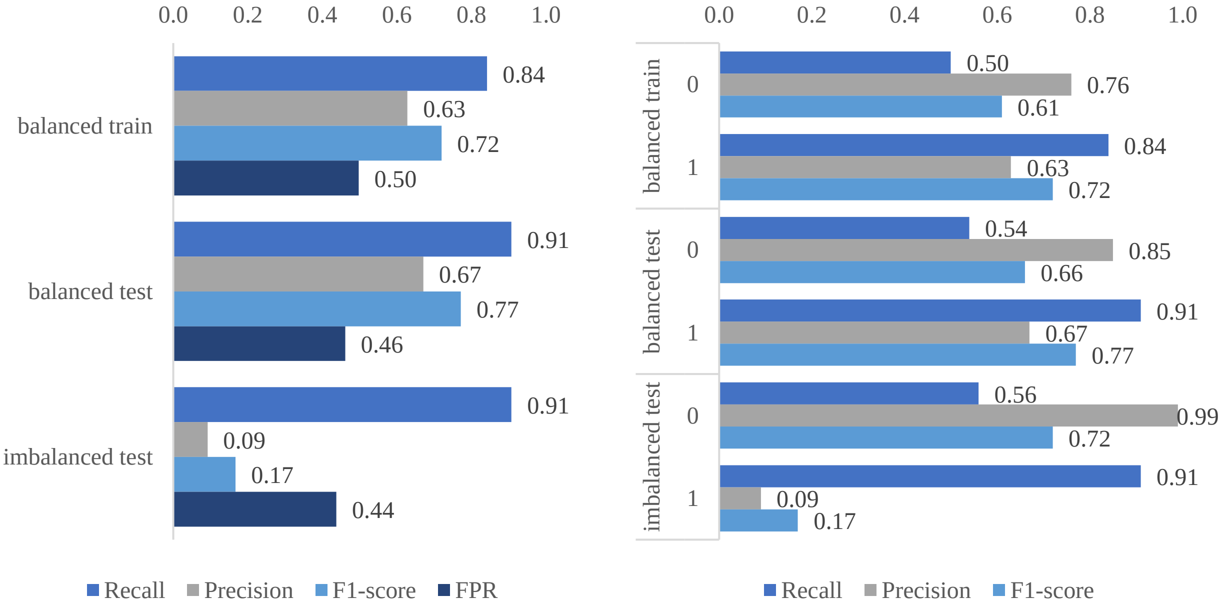


Figure 10. Performance of *svm_wc* (left) and per class (right) tenfold cross-validation and test results

The models' varying robustness and performance consistency can be attributed to the difference in training data size. *svm_p2* used a training set almost three times smaller than *svm_wc* (i.e., n_{p2}=157 and n_{wc}=418), making it more susceptible to overfitting and less capable of learning from the available data.

Furthermore, on the imbalanced test set, both *svm_p2* and *svm_wc* achieved a high recall but a very low precision, resulting to a low f1-score of 27% and 17% respectively (see Figure 9 and 10 left under imbalanced test). Both models showed poor performance in predicting positive instances, which is likely due to the fewer number of positive instances compared to negative instances (see Figure 9 and10 right under imbalanced test). This finding emphasizes the need to consider class imbalance in model evaluation to improve performance.

Overall, *svm_wc* showed promising performance on the balanced training and test sets. Due to the robustness of the model *svm_wc* compared to *svm_p2*, the focus of the discussion on the next sections is on the results of *svm_wc*.

For the next analysis, the weights of the bigrams assigned by the model *svm_wc* during the training phase were examined. The further a data point is from the hyperplane, the more positive the weight is for the positive class (*wc*=1) or the more negative the weight is for the negative class (*wc*=0).

Figure 11 displays the top 30 bigrams for *svm_wc* that ranked the highest to identify a merger case to be 'approved unconditionally' (blue bars) or 'approved with conditions' (red bars). The top 100 features by importance per class are available in Appendix B Table B.5.

Certain keywords were found to be merger-specific, identifying specific markets and companies. Keywords that predicted cases 'approved unconditionally' include *abn amro*, *fortis abn*, *comp fortis*, *Deutsche bank* and *investment banking*, which appear to be linked to the banking market. Some were names of firms based on a particular case (i.e., Fortis and ABN AMRO).[18] Moreover, features like *life insurance, non life,* and *insurance product* most likely pertain to the insurance industry, whereas the bigrams *gas oil, low voltage,* and *gas sector* are associated with the gas sector.

Meanwhile, keywords that predicted cases 'approved with conditions' like *additional value, million achieves,* and *profitable business* seem to be associated with

[18] Case No. M.4844 - FORTIS / ABN AMRO ASSETS

large profitable transactions. Certain keywords also appear to be market-specific, for instance, related to health and pharmaceutical sector (e.g., *generic pharmaceutical* and *health business*), while some may be specific geographic markets (e.g., *Asian country* and *Hong Kong*).

Some "meaningless" terms like *ii* and *ha* were also detected in the bigrams by *svm_wc*, due to the model's simplicity and the absence of additional filtering methods to eliminate irrelevant information. Nevertheless, the test results of *svm_wc* were comparable to the cross-validation results, indicating that the model was performing well despite the use of only basic textual features.

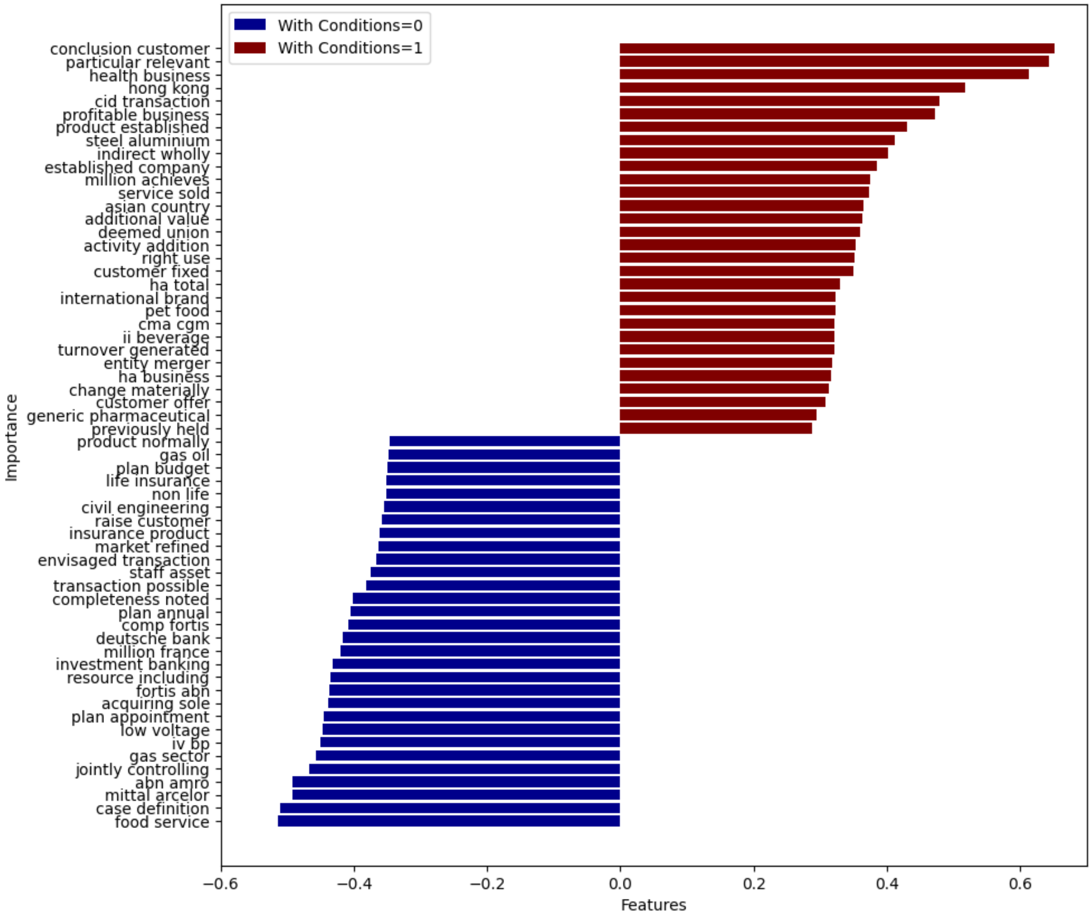


Figure 11. Top 30 features by importance per class of *svm_wc*

5.1.3. Discussion

Although the model *svm_p2* was found to be unreliable due to small training data, *svm_wc* was determined to be a robust and high-performing model, achieving 84% recall and 63% precision under cross-validation.

The confusion matrix in Table 9 summarizes the tenfold cross-validation results for 418 cases equally split between classes. Out of 209 cases 'approved unconditionally', 105 were identified correctly and 104 were identified incorrectly. Out of 209 'approved with conditions', 176 cases were classified correctly and 33 cases were classified incorrectly. This suggests that the model performed better at predicting cases 'approved with conditions' than cases 'approved unconditionally'.

Table 9. Confusion matrix of *svm_wc* tenfold cross-validation

		Actual	
		Approved unconditionally	Approved with conditions
Predicted	Approved unconditionally	105 (TN)	33 (FN)
	Approved with conditions	104 (FP)	176 (TP)

Note: TN=true negatives, FN=false negatives, FP=false positives, TP=true positives

The prediction errors of *svm_wc* were manually reviewed based on the economic activities involved in the merger. Several true negatives included cases in electricity generation, fuel production, and financial and insurance services. False negatives were found in cases related to electric power generation and electronic equipment manufacturing. False positives were identified in some transactions concerned in telecommunications and manufacture of pharmaceutical products. True positives were also found in telecommunications and manufacture of pharmaceutical products, medical and dental supplies, and chemicals.

Table 10 displays the most frequent unigrams and bigrams from the NACE description of cases, categorized into the four types of outcomes shown in the confusion matrix. Sectors in bold correspond to market-specific keywords in Figure 11. For instance, *generic pharmaceutical* and *health business* were important features for predicting cases 'approved with conditions', which are related to the sectors in bold under TP. Similarly, keywords for predicting cases 'approved unconditionally' like *life insurance, non life,* and *insurance product* may be related to pension funding under TN. While the relationship of the NACE description and keywords is rather speculative, it highlights that certain features relate to the economic activities and relevant markets involved in the merger.[19]

[19] Pension funding: *life insurance, non life,* and *insurance product*;
Gaseous fuels and petroleum products: *gas oil, low voltage,* and *gas sector*;

However, it is important to note that merger review not only depends on these variables but also other factors like competitive constraints, market entry, efficiencies, and more.

Table 10. Most frequent unigrams and bigrams of NACE descriptions by TN, FN, FP, and TP

		Actual	
		Approved unconditionally	Approved with conditions
Predicted	Approved unconditionally	(TN)	(FN)
		motor vehicles	electricity production
		chemical products	motor vehicles
		pension funding	**hospital activities**
		gaseous fuels	**pension funding**
		petroleum products	**pharmaceutical preparations**
	Approved with conditions	(FP)	(TP)
		telecommunications	telecommunications
		chemical products	**pharmaceutical products**
		motor vehicles	**pharmaceutical preparations**
		computer programming	chemical products
		programming consultancy	**dental instruments**

Note: Bold texts indicates that the sectors are related to certain features of *svm_wc*.

Hospital activities, pharmaceutical products and preparation, dental instruments: *generic pharmaceutical* and *health business*

5.2. Experiment 2: Text classification of *wc* using single section text

5.2.1. Set-up

In Experiment 2, the model *svm_wc* was trained and evaluated separately on text from *Parties and Operation (po), Concentration & Dimension (cd), Market Definition (md),* and *Competitive Assessment (ca)*, namely *svm_wc_po, svm_wc_cd, svm_wc_md,* and *svm_wc_ca*, respectively. The models were trained using the same set of parameters from Table 8, with tenfold cross-validation and 5000 bigram features, and evaluated on the test set.

As some cases were missing certain sections, the training data size for Experiment 2 was reduced to 38-132 cases per section, compared to the 418 cases in Experiment 1. Specifically, n_{wc_po}=345, n_{wc_cd}=380, n_{wc_md}=286, and n_{wc_ca}=332 with each having a balanced number between positive and negative classes.

5.2.2. Results

Under tenfold cross-validation, the recall was higher when the model was trained on *Market Definition* and *Competitive Assessment* than the other sections (Figure 12). The higher performance was expected as the models were trained with the most informative sections, including the substantial analysis and economic arguments necessary to arrive at a decision. Appendix D contains the results for all four models.

In the balanced test set, *svm_wc_md* and *svm_wc_ca* also resulted to a high recall, lower precision, and a high FPR (Figure 13 left). Both models were also better at predicting cases 'approved with conditions' than cases 'approved unconditionally' (Appendix D Table D.2 and D.3).

Likewise, the top features of both models were found to be market-specific. Keywords of *svm_wc_md* include *electronic component, supply automotive,* and *distribution electricity*, while *svm_wc_ca*'s features include *water treatment, electricity generation, fuel oil,* and *supply electricity*.

5.2.3. Discussion

Experiment 2 shows that a targeted feature selection led to better performance. Under cross-validation, *svm_wc_md* and *svm_wc_ca* achieved a recall of 94% and 98%, respectively, with precision of 58%, higher than Experiment 1's *svm_wc* by 10 to 14 percentage points in recall and lower by 5 percentage points in precision.

Balance between recall and precision is crucial to select the best model for this study since a higher recall leads to more false positives, which also have negative implications for predicting merger decision outcomes.

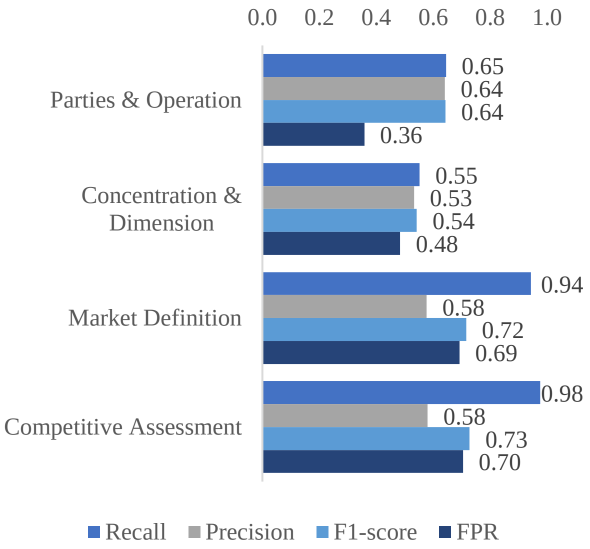


Figure 12. Performance of *svm_wc_po*, *svm_wc_cd*, *svm_wc_md*, and *svm_wc_ca* tenfold cross-validation

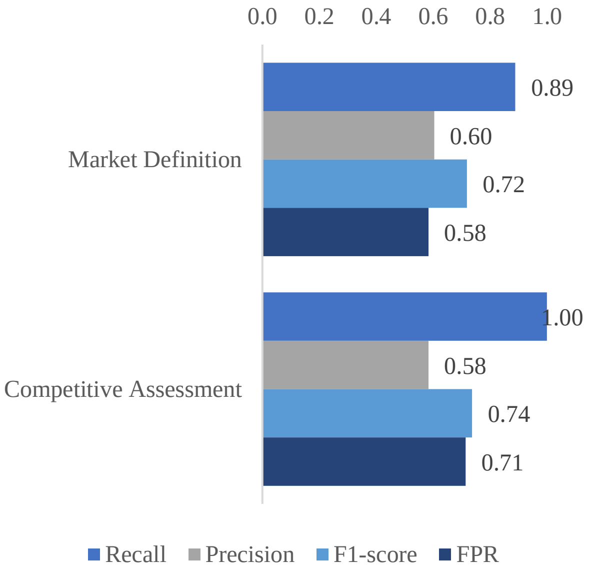


Figure 13. Performance of *svm_wc_md* and *svm_wc_ca* test results

6. Policy Implications

The models achieved the desired performance with higher recall than precision and a relatively high FPR. Experiment 1's *svm_wc* showed a high performance, achieving 84% recall, 63% precision, and 50% FPR under cross-validation. In Experiment 2, both *svm_wc_md* and *svm_wc_ca* outperformed the previous model, achieving a recall of 94% and 98%, respectively, and 58% precision and about 70% FPR for both.

To understand the relevance of the attained model performance, two potential events, event A and event B, in a merger review must be considered. Event A involves approving a case unconditionally despite the presence of serious anticompetitive effects, while Event B involves approving a case with conditions despite little to no anticompetitive effects. In the context of this study, the costs and benefits of each event must be assessed.

The cost of Event A is the potential harm to competition, especially to consumers, in the form of higher prices, decreased quality of goods and services, or reduced innovation. The benefits of Event A include potential cost savings for the merged entity and fewer regulatory burdens.

Meanwhile, the cost of Event B involves resources for monitoring compliance with the conditions of the merged entity and limiting potential efficiencies leading to increased compliance costs for firms and delays in market entry. However, the benefits of Event B include preventing potential competition harm by imposing conditions on the merged entity, such as divestitures or restrictions on certain business practices.

In this study, Event B was preferred over Event A. This is why a high recall was chosen as a model evaluation metric, due to the severe consequences of overlooking problematic transactions. However, it is still important to note that a high false positive rate can also have negative consequences, as highlighted in the costs of Event B.

Hence, while a high recall is important for identifying cases with serious anticompetitive effects, it should be balanced with a reasonable precision to avoid imposing unnecessary conditions on cases that do not have anticompetitive effects. Given these considerations, the Experiment 1's *svm*_wc is deemed as the best model having a better balance between recall and precision.

The model *svm_wc* was not only robust but also picked up certain features based on specific economic activities and markets involved in the proposed mergers. It was able to accurately predict the outcome of similar cases in some sectors but it also had prediction errors in certain sectors.

The performance of *svm_wc* may be affected by several factors including the quality and completeness of the training data, relevance of case similarity, and complexity of merger decisions. Therefore, ML models like *svm_wc* should only be used as a supplementary tool to expert judgment to ensure a thorough analysis of all relevant factors for making the final decision.

Nevertheless, this study serves as an initial step in exploring the applications of data science in legal analysis and antitrust review, which can be extended to developing a tool that identifies hard-to-detect theories of harm and recommends remedies from past merger cases to future cases to promote competition.

7. Conclusion and Future Work

This study demonstrates that a text-based approach with NLP and ML can be used to predict EC merger decision outcomes using extracted text from official merger decisions. An SVM linear classifier was found to be the best model for predicting mergers that were approved with or without conditions. This model yielded a high recall of 84% to identify cases with serious anticompetitive effects and a relatively lower precision of 63% to avoid imposing unnecessary conditions on cases that do not have anticompetitive effects. It also used keywords on relevant markets, indicating its ability to capture classification nuances. Practical use requires a cost-benefit analysis to determine the optimal trade-off between recall and precision based on EC's specific objectives and priorities.

Overall, this study provides useful insights into the potential of NLP and ML techniques for predicting antitrust decisions, but it also underscores the model's limitations and challenges in practical applications. Expert judgment and human analysis should always be considered to ensure a comprehensive and reliable decision-making process.

Future work can improve on the data size to avoid overfitting and ensure the model generalizes well to new data by including cases reviewed before 2004 under the "old" EU Merger Control Regulation and decisions from other jurisdictions like the UK Competition and Markets Authority. To address class imbalance and avoid biased predictions, data augmentation can also be explored to increase training data.

The text parsing technique can also be enhanced to ensure more accurate section-by-section parsing and remove irrelevant information. Data from other sources like company reports and press releases about proposed mergers can be extracted and merged with the current dataset to explore relationships between variables.

Furthermore, it can be valuable to consider the time variable when splitting the training and test sets. This involves using older cases for training and more recent cases for testing to examine the influence of precedents on decision-making and outcomes.

Finally, other steps in the text classification pipeline can be modified by using word embeddings to consider semantics, and trying different ML algorithms like Tree-based models and XGBoost. Using deep learning algorithms can likewise be explored, including multilayer perceptron, recurrent neural networks, and transformers. Pre-trained models like BERT can be fine-tuned, however, this would be a black box approach and would not provide information on the importance of textual features.

Bibliography

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *PeerJ Computer Science*, *2*, e93. https://doi.org/10.7717/peerj-cs.93

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc.*

Council Regulation (EC) No 139/2004. Retrieved January 31, 2023, from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32004R0139

Council Regulation (EEC) No 4064/89 (the "old" EU Merger Regulation). Retrieved January 31, 2023, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31989R4064

Danilak, M. (2021). *langdetect*. Retrieved May 1, 2023, from https://pypi.org/project/langdetect/

European Commission. (2021). Report on Competition Policy 2021. Retrieved January 31, 2023, from https://competition-policy.ec.europa.eu/publications/annual-reports_en

European Commission. (2013). Merger control procedures. Retrieved January 31, 2023, from https://competition-policy.ec.europa.eu/system/files/2021-02/merger_control_procedures_en.pdf

European Commission. (2004). Guidelines on the assessment of horizontal mergers Retrieved January 31, 2023, from https://eur-lex.europa.eu/EN/legal-content/summary/guidelines-on-the-assessment-of-horizontal-mergers.html

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). https://www.nature.com/articles/s41586-020-2649-2

Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*. https://doi.org/10.1086/688176

Jiang, T. (2021). Using Machine Learning to Analyze Merger Activity. *Frontiers in Applied Mathematics and Statistics*, *7*, 649501. https://doi.org/10.3389/fams.2021.649501

Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management*, *56*(5), 1618–1632. https://doi.org/10.1016/j.ipm.2019.05.003

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1-5. Retrieved May 1, 2023, from http://jmlr.org/papers/v18/16-365.html

McKinney, W., & others. (2010). Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

Moriarty, R., Ly, H., Lan, E., & McIntosh, S. K. (2019). Deal or No Deal: Predicting Mergers and Acquisitions at Scale. *2019 IEEE International Conference on Big Data (Big Data)*, 5552–5558. https://doi.org/10.1109/BigData47090.2019.9006015

Morzenti, G. (2022). Antitrust Policy and Innovation. *University of Bocconi Department of Economics*, *Working Paper*.

pdfplumber. (2023). Retrieved May 1, 2023, from https://github.com/jsvine/pdfplumber

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Python Software Foundation. (2021). Regular expression operations. Retrieved Retrieved May 1, 2023, from https://docs.python.org/3/library/re.html

Ragab, A., El Koujok, M., Ghezzaz, H., & Amazouz, M. (2021). Fault diagnosis in industrial processes based on predictive and descriptive machine learning methods. *Applications of Artificial Intelligence in Process Systems Engineering*, 207-254. https://doi.org/10.1016/B978-0-12-821092-5.00002-4

Routledge, B. R., Sacchetto, S., & Smith, N. A. (2017). Predicting Merger Targets and Acquirers from Text. *Carnegie Mellon University*, *Working Paper*.

Virtucio, M. B. L., Aborot, J. A., Abonita, J. K. C., Avinante, R. S., Copino, R. J. B., Neverida, M. P., Osiana, V. O., Peramo, E. C., Syjuco, J. G., & Tan, G. B. A. (2018). Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 130–135. https://doi.org/10.1109/COMPSAC.2018.10348

Washington State Department of Revenue. (n.d.). SIC and NAICS codes. Retrieved January 31, 2023, from https://dor.wa.gov/about/statistics-reports/sic-and-naics-codes#:~:text=QBR%20NAICS%20data%3F-What%20are%20SIC%20and%20NAICS%20codes%3F,primarily%20engaged%20in%20food%20manufacturing

White & Case. (2022, January 10). Shining a light on the massive global surge in merger control filings. White & Case LLP. Retrieved January 31, 2023, from https://www.whitecase.com/insight-our-thinking/shining-light-massive-global-surge-merger-control-filings

Appendix A. Feature Matrix Generation

Using *tf-idf* scores of *n-grams* or sequences of *n* words provides a way to capture the most important and relevant terms in a document relative to the entire corpus.

In this study, the maximum word sequence was limited to three since longer *n-grams* are less likely to occur. The *n-grams* were categorized into three types: unigrams, bigrams, and trigrams, which represent single words, sequences of two words, and sequences of three words, respectively. For instance, consider the following sentence from the *Parties* section of the case cited in Figure 4.

Glatfelter is a New York stock market listed manufacturer in the "specialty papers" and "composite fibres" business areas and has production sites in the USA, the Philippines, France and Germany.

Since the raw text was pre-processed (i.e., lowercasing, lemmatization, and removal of punctuations, numeric, etc.), the bigrams of the sentence would consist of:

glatfelter is, is a, a new, new york, york stock, stock market, market listed, listed manufacture, manufacture in, in the, the specialty, specialty papers, papers and, and composite, composite fibres, fibres business, areas and, and has, has production, production sites, sites in, in the, the usa, usa the, the philippines, philippines france, france and, and Germany

The *tf-idf* was then calculated as:

$$tfidf(d,t) = tf(t) * idf(d,t), \text{ where}$$

$$tf = \frac{\text{no. of times term } t \text{ appears in document } d}{\text{no. of terms in document } d}$$

$$idf(d,t) = \log\left(\frac{n}{df(d,t)}\right) + 1,$$

$$n = \text{no. of documents in the corpus},$$

$$df(d,t) = \text{no. of documents in the corpus that contain term } t.$$

For example, a corpus contains 2,000 documents and a 500-word document contains the bigram "stock market" five times which appears in 20 documents. The term frequency (*tf*) is $\frac{5}{500} = 0.01$ and the inverse document frequency (*idf*) is $\log\left(\frac{2{,}000}{20}\right) + 1 = 3$. Then, the *tf-idf* score is. $0.01 * 3 = 0.03$.

Appendix B. Experiment 1 Results

Table B.1. Distribution of merger decisions by section

Section	No. of cases
Parties & Operation	1,442
Concentration & Dimension	1,538
Market Definition	1,051
Competitive Assessment	1,331
Total	**5,362**

Table B.2. Recall, precision, f1-score, and FPR of *svm_p2* and *svm_wc* tenfold cross-validation and test results

Model	Set	n	Recall	Precision	F1-score	FPR
Reviewed for phase 2						
svm_p2	balanced train	157	0.83	0.90	0.87	0.09
	balanced test	40	0.80	0.94	0.86	0.05
	imbalanced test	1,426	0.80	0.16	0.27	0.06
Approved with conditions						
svm_wc	balanced train	418	0.84	0.63	0.72	0.50
	balanced test	106	0.91	0.67	0.77	0.46
	imbalanced test	1,158	0.91	0.09	0.17	0.44

Table B.3. Recall, precision, and f1-score of *svm_p2* and *svm_wc* per class of tenfold cross-validation and test results

Model	Set	Class	n	Recall	Precision	F1-score
Reviewed for phase 2						
svm_p2	balanced train	0	79	0.91	0.85	0.88
		1	78	0.83	0.9	0.87
	balanced test	0	20	0.95	0.83	0.88
		1	20	0.8	0.94	0.86
	imbalanced test	0	1,406	0.94	1.00	0.97
		1	20	0.80	0.16	0.27
Approved with conditions						
svm_wc	balanced train	0	209	0.50	0.76	0.61
		1	209	0.84	0.63	0.72
	balanced test	0	52	0.54	0.85	0.66
		1	54	0.91	0.67	0.77
	imbalanced test	0	1,104	0.56	0.99	0.72
		1	54	0.91	0.09	0.17

Table B.4. Confusion matrix of *svm_p2* and *svm_wc* of tenfold cross-validation and test results

Model	**Set**	**n**	**TN**	**FP**	**FN**	**TP**
Reviewed for phase 2						
svm_p2	balanced train	157	72	7	13	65
	balanced test	40	19	1	4	16
	imbalanced test	1,426	1,323	83	4	16
Approved with conditions						
svm_wc	balanced train	418	105	104	33	176
	balanced test	106	28	24	5	49
	imbalanced test	1,158	621	483	5	49

Table B.5. Top 100 features by importance per class of *svm_wc*

	***wc*=1**		***wc*=0**	
	feature	**weight**	**feature**	**weight**
1	conclusion customer	0.65	food service	-0.51
2	particular relevant	0.64	case definition	-0.51
3	health business	0.61	mittal arcelor	-0.49
4	hong kong	0.52	abn amro	-0.49
5	cid transaction	0.48	jointly controlling	-0.47
6	profitable business	0.47	gas sector	-0.46
7	product established	0.43	iv bp	-0.45
8	steel aluminium	0.41	low voltage	-0.45
9	indirect wholly	0.40	plan appointment	-0.45
10	established company	0.39	acquiring sole	-0.44
11	million achieves	0.37	fortis abn	-0.44
12	service sold	0.37	resource including	-0.43
13	asian country	0.36	investment banking	-0.43
14	additional value	0.36	million france	-0.42
15	deemed union	0.36	deutsche bank	-0.42
16	activity addition	0.35	comp fortis	-0.41
17	right use	0.35	plan annual	-0.41
18	customer fixed	0.35	completeness noted	-0.40
19	ha total	0.33	transaction possible	-0.38
20	international brand	0.32	staff asset	-0.37
21	pet food	0.32	envisaged transaction	-0.37
22	cma cgm	0.32	market refined	-0.36
23	ii beverage	0.32	insurance product	-0.36
24	turnover generated	0.32	raise customer	-0.36
25	entity merger	0.32	civil engineering	-0.35
51	account difference	0.20	professional customer	-0.30
52	euro million	0.20	overlap proposed	-0.30
53	market provides	0.20	channel wa	-0.30
54	commitment annexed	0.20	located region	-0.29
55	point national	0.20	oil refinery	-0.29
56	remains important	0.19	service vertical	-0.29
57	constitute obligation	0.19	national national	-0.29
58	implementing step	0.19	wide regional	-0.29
59	data protection	0.19	product life	-0.28
60	proposed divest	0.19	comp exxon	-0.28
61	eea consumption	0.19	transaction jv	-0.28
62	gathered course	0.19	states aggregate	-0.28
63	clear picture	0.18	million proposed	-0.28
64	france addition	0.18	financial risk	-0.28
65	kingdom germany	0.18	jv business	-0.28
66	partially overlapping	0.18	used vehicle	-0.28
67	concern implementing	0.18	area application	-0.28
68	constitute condition	0.18	commission envisaged	-0.28
69	meaningful market	0.18	netherlands according	-0.27
70	arab emirates	0.18	owned indirect	-0.27
71	united arab	0.18	dow chemical	-0.27
72	current shareholder	0.18	product sub	-0.27
73	commitment set	0.18	building material	-0.27
74	commission party	0.18	technology ii	-0.27
75	case decision	0.18	table proposed	-0.27

	***wc*=1**		***wc*=0**	
26	ha business	0.32	life insurance	-0.35
27	change materially	0.31	non life	-0.35
28	customer offer	0.31	plan budget	-0.35
29	generic pharmaceutical	0.30	gas oil	-0.35
30	previously held	0.29	product normally	-0.35
31	spanish company	0.29	vertical aspect	-0.34
32	distribution retail	0.28	right regard	-0.33
33	product pipeline	0.28	approximately worldwide	-0.33
34	commission mean	0.28	service operation	-0.33
35	relevant member	0.27	change regard	-0.33
36	case time	0.27	cid network	-0.33
37	relating sale	0.25	investment management	-0.32
38	operates production	0.25	approx eur	-0.32
39	foreclose customer	0.24	free choose	-0.32
40	power allow	0.24	possibly wider	-0.32
41	set annex	0.23	venture ha	-0.32
42	manufacturing operation	0.23	transaction agreement	-0.32
43	leading brand	0.23	turnover period	-0.32
44	require approval	0.22	share target	-0.31
45	identified party	0.22	equity fund	-0.31
46	price related	0.21	chairman board	-0.31
47	service account	0.21	customer primarily	-0.30
48	strong focus	0.21	eu turnover	-0.30
49	obligation concern	0.20	director appointed	-0.30
50	spain belgium	0.20	totalfina elf	-0.30
76	submitted form	0.18	market steel	-0.27
77	purchase non	0.17	considers operation	-0.27
78	online service	0.17	based individual	-0.27
79	regulation paragraph	0.17	variety industry	-0.27
80	divided main	0.17	financial leasing	-0.27
81	place order	0.17	transaction precise	-0.27
82	basis demand	0.17	ii upstream	-0.27
83	material commission	0.17	bain capital	-0.26
84	point separate	0.16	national left	-0.26
85	using alternative	0.16	affected vertical	-0.26
86	transaction deemed	0.16	group parties	-0.26
87	discussed separately	0.16	wa segmented	-0.26
88	final commitment	0.16	operation concern	-0.26
89	brand party	0.16	upstream retail	-0.26
90	market produce	0.16	involved production	-0.26
91	belgium cyprus	0.16	market flat	-0.26
92	venture jv	0.16	product horizontal	-0.26
93	capital increase	0.16	party additionally	-0.25
94	case analysis	0.15	eur basis	-0.25
95	possible competition	0.15	possible national	-0.25
96	used prevent	0.15	type used	-0.25
97	economic area	0.15	commission reasoned	-0.25
98	oj february	0.15	steel market	-0.25
99	national reimbursement	0.15	right general	-0.25
100	late stage	0.15	flat product	-0.25

Appendix C. Experiment 1 Robustness Results

Table C.1. Robustness check of *svm_p2* and *svm_wc* using different random seed and state

Model	Set	n	Recall	Precision	F1-score	FPR
Reviewed for phase 2						
svm_p2_28	balanced train	158	0.99	0.58	0.73	0.70
	balanced test	41	0.90	0.59	0.72	0.65
	imbalanced test	1,427	0.90	0.02	0.04	0.74
svm_p2_70	balanced train	157	0.99	0.57	0.73	0.72
	balanced test	40	1.00	0.63	0.77	0.60
	imbalanced test	1,426	1.00	0.02	0.04	0.69
Approved with conditions						
svm_wc_28	balanced train	417	0.83	0.62	0.71	0.50
	balanced test	105	0.85	0.69	0.76	0.38
	imbalanced test	1,157	0.85	0.08	0.15	0.44
svm_wc_70	balanced train	417	0.92	0.66	0.77	0.47
	balanced test	105	0.90	0.67	0.77	0.43
	imbalanced test	1,157	0.90	0.08	0.14	0.50

Table C.2. Robustness check of *svm_p2* and *svm_wc* per class using different random seed and state

Model	Set	Class	n	Recall	Precision	F1-score
Reviewed for phase 2						
svm_p2_28	balanced train	0	80	0.30	0.96	0.46
		1	70	0.99	0.58	0.73
	balanced test	0	20	0.35	0.78	0.48
		1	21	0.90	0.59	0.72
	imbalanced test	0	1,406	0.26	0.99	0.42
		1	21	0.90	0.02	0.04
svm_p2_70	balanced train	0	79	0.28	0.96	0.43
		1	78	0.99	0.57	0.73
	balanced test	0	20	0.40	1.00	0.57
		1	20	1.00	0.62	0.77
	imbalanced test	0	1,406	0.31	1.00	0.48
		1	20	1.00	0.02	0.04
Approved with conditions						
svm_wc_28	balanced train	0	208	0.50	0.74	0.60
		1	209	0.83	0.62	0.71
	balanced test	0	53	0.62	0.80	0.70
		1	52	0.85	0.69	0.76
	imbalanced test	0	1,105	0.56	0.99	0.99
		1	52	0.85	0.15	0.08

Model	Set	Class	n	Recall	Precision	F1-score
svm_wc_70	balanced train	0	208	0.53	0.87	0.66
		1	209	0.92	0.66	0.77
	balanced test	0	53	0.57	0.86	0.68
		1	52	0.90	0.67	0.77
	imbalanced test	0	1105	0.50	0.99	0.66
		1	52	0.90	0.08	0.14

Table C.3. Robustness check of confusion matrix of *svm_p2* and *svm_wc* using different random seed and state

Model	Set	n	TN	FP	FN	TP
Reviewed for phase 2						
svm_p2_28	balanced train	158	24	56	1	77
	balanced test	41	7	13	2	19
	imbalanced test	1,427	371	1035	2	19
svm_p2_70	balanced train	157	22	57	1	77
	balanced test	40	8	12	0	20
	imbalanced test	1,426	442	964	0	20
Approved with conditions						
svm_p2_28	balanced train	417	104	104	36	173
	balanced test	105	33	20	8	44
	imbalanced test	1,157	617	488	8	44
svm_p2_70	balanced train	417	111	97	17	192
	balanced test	105	30	23	5	47
	imbalanced test	1,157	550	555	5	47

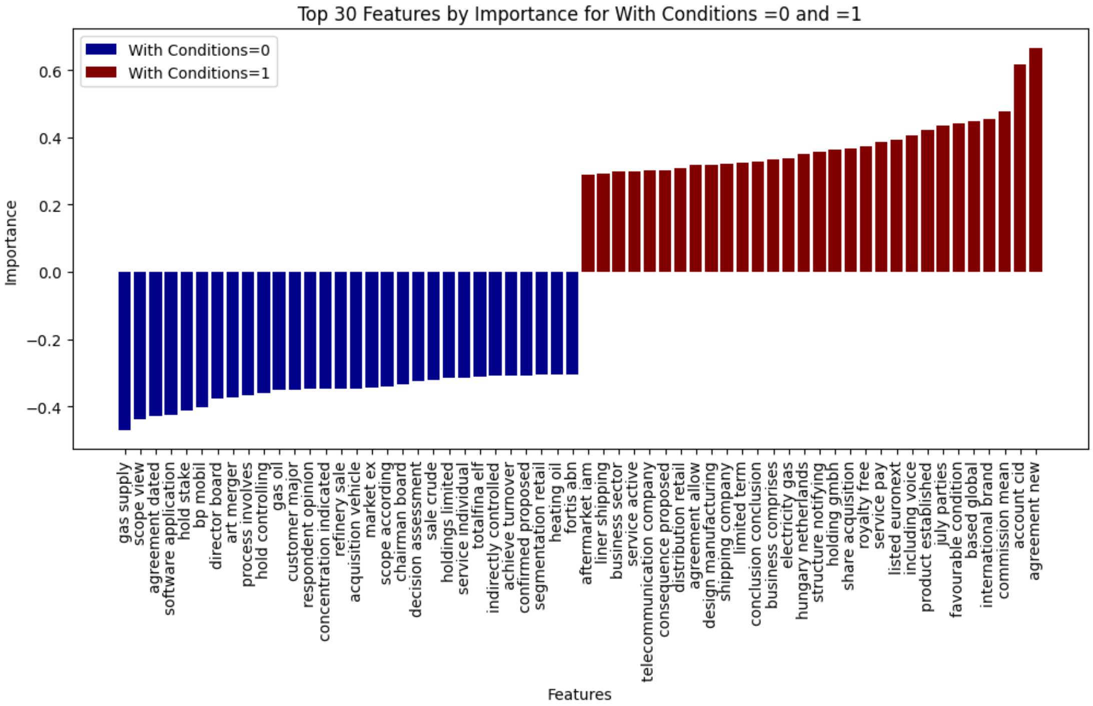


Figure C.1. Top 30 features by importance of *svm_wc* per class using random seed and random state=28

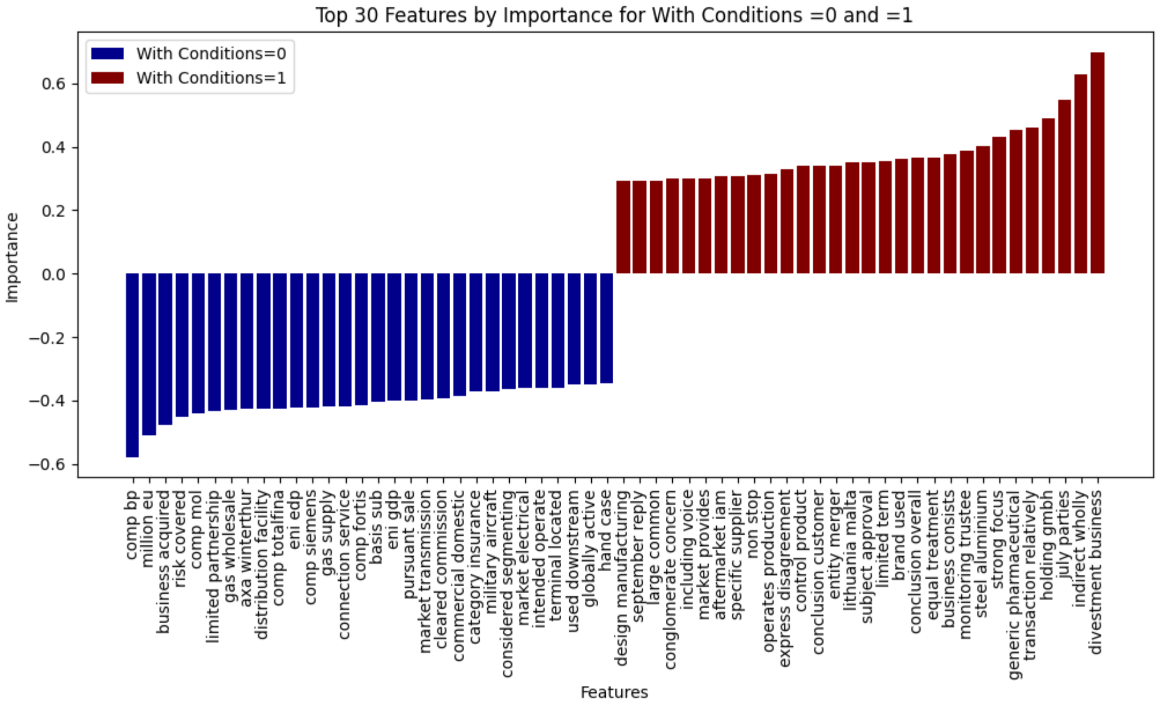


Figure C.2. Top 30 features by importance of *svm_wc* per class using random seed and random state=70

Appendix D. Experiment 2 Results

Table D.1. Recall, precision, f1-score, and FPR of *svm_wc* tenfold cross-validation and test results by section

Model	Set	n	Recall	Precision	F1-score	FPR
Parties & Operation						
svm_wc_po	balanced train	345	0.65	0.64	0.64	0.36
	balanced test	87	0.66	0.66	0.66	0.35
	imbalanced test	1,090	0.66	0.07	0.13	0.37
Concentration & Dimension						
svm_wc_cd	balanced train	380	0.55	0.53	0.54	0.48
	balanced test	96	0.33	0.47	0.39	0.38
	imbalanced test	1,152	0.33	0.05	0.08	0.29
Market Definition						
svm_wc_md	balanced train	286	0.94	0.58	0.72	0.69
	balanced test	72	0.89	0.60	0.72	0.58
	imbalanced test	758	0.89	0.06	0.12	0.65
Competitive Assessment						
svm_wc_ca	balanced train	332	0.98	0.58	0.73	0.70
	balanced test	84	1.00	0.58	0.74	0.71
	imbalanced test	990	1.00	0.07	0.12	0.64

Table D.2. Recall, precision, and f1-score of *svm_wc* per class of tenfold cross-validation and test results by section

Model	Set	Class	n	Recall	Precision	F1-score
Parties & Operation						
svm_wc_po	balanced train	0	173	0.64	0.65	0.64
		1	172	0.65	0.64	0.64
	balanced test	0	43	0.65	0.65	0.65
		1	44	0.66	0.66	0.66
	imbalanced test	0	1046	0.63	0.98	0.77
		1	44	0.66	0.07	0.13
Concentration & Dimension						
svm_wc_cd	balanced train	0	190	0.52	0.54	0.53
		1	190	0.55	0.53	0.54
	balanced test	0	48	0.62	0.48	0.55
		1	48	0.33	0.47	0.39
	imbalanced test	0	1104	0.71	0.96	0.82
		1	48	0.33	0.05	0.08
Market Definition						
svm_wc_md	balanced train	0	143	0.31	0.85	0.45
		1	143	0.94	0.58	0.72
	balanced test	0	36	0.42	0.79	0.55

Model	**Set**	**Class**	**n**	**Recall**	**Precision**	**F1-score**
		1	36	0.89	0.6	0.72
	imbalanced test	0	732	0.35	0.98	0.52
		1	36	0.89	0.06	0.12
Competitive Assessment						
svm_wc_ca	balanced train	0	166	0.3	0.92	0.45
		1	166	0.98	0.58	0.75
	balanced test	0	42	0.29	1.00	0.44
		1	42	1.00	0.58	0.74
	imbalanced test	0	948	0.36	1.00	0.53
		1	42	1.00	0.07	0.12

Table D.3. Confusion matrix of *svm_wc* of tenfold cross-validation and test results by section

Model	**Set**	**n**	**TN**	**FP**	**FN**	**TP**
Parties & Operation						
svm_wc_po	balanced train	345	111	62	61	111
	balanced test	87	28	15	15	29
	imbalanced test	1,090	659	387	15	29
Concentration & Dimension						
svm_wc_cd	balanced train	380	98	92	85	105
	balanced test	96	30	18	32	16
	imbalanced test	1,152	786	318	32	16
Market Definition						
svm_wc_md	balanced train	286	44	99	8	135
	balanced test	72	15	21	4	32
	imbalanced test	758	253	469	4	32
Competitive Assessment						
svm_wc_ca	balanced train	332	49	117	4	162
	balanced test	84	12	30	0	42
	imbalanced test	990	345	603	0	42

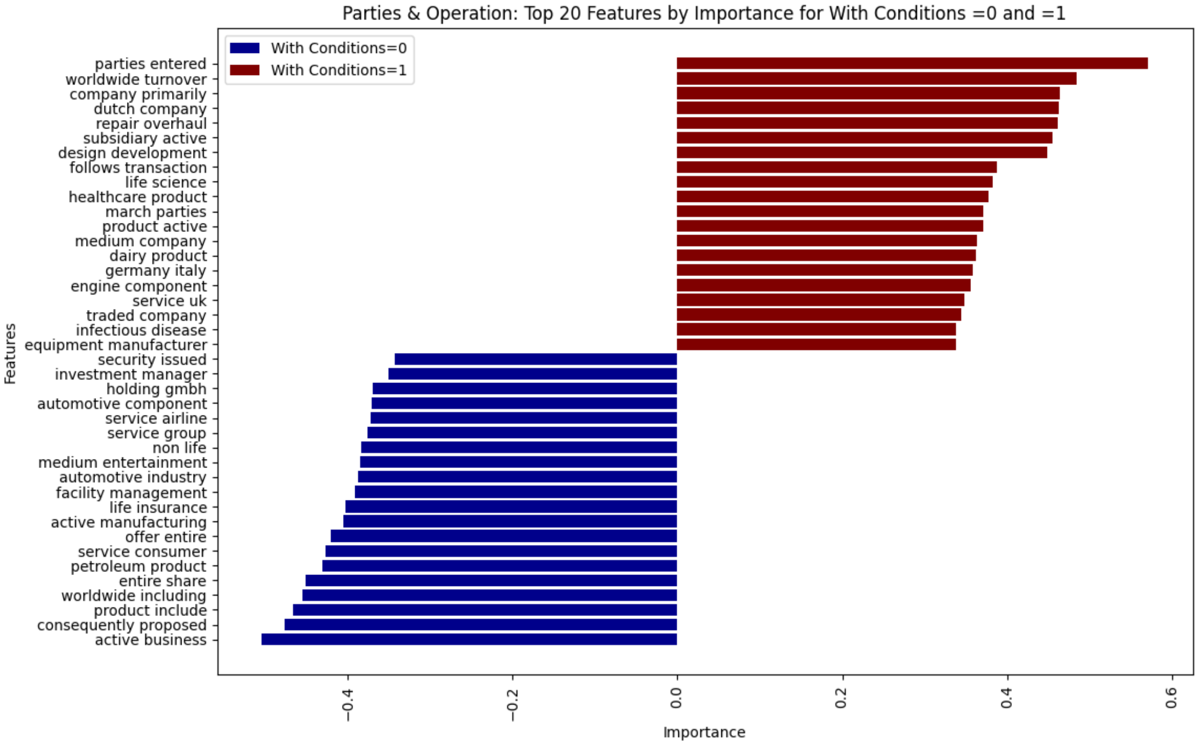


Figure D.1. Top 20 features by importance of *svm_wc_po* per class

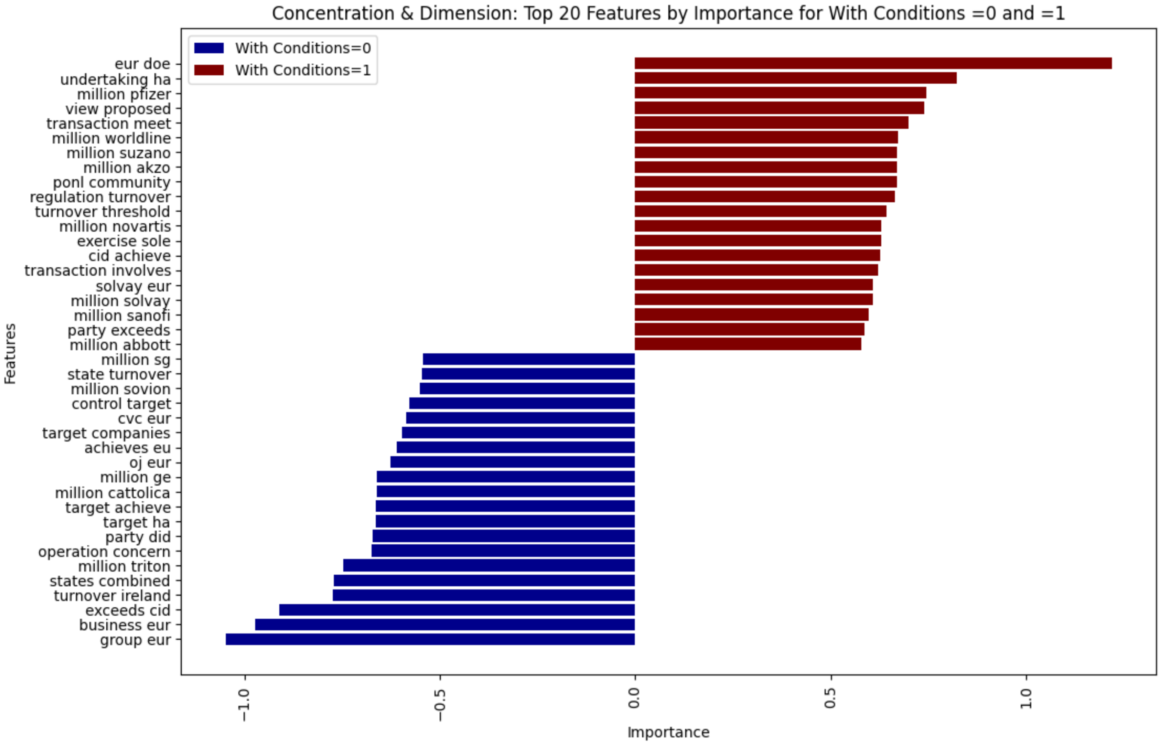


Figure D.2. Top 20 features by importance of *svm_wc_cd* per class

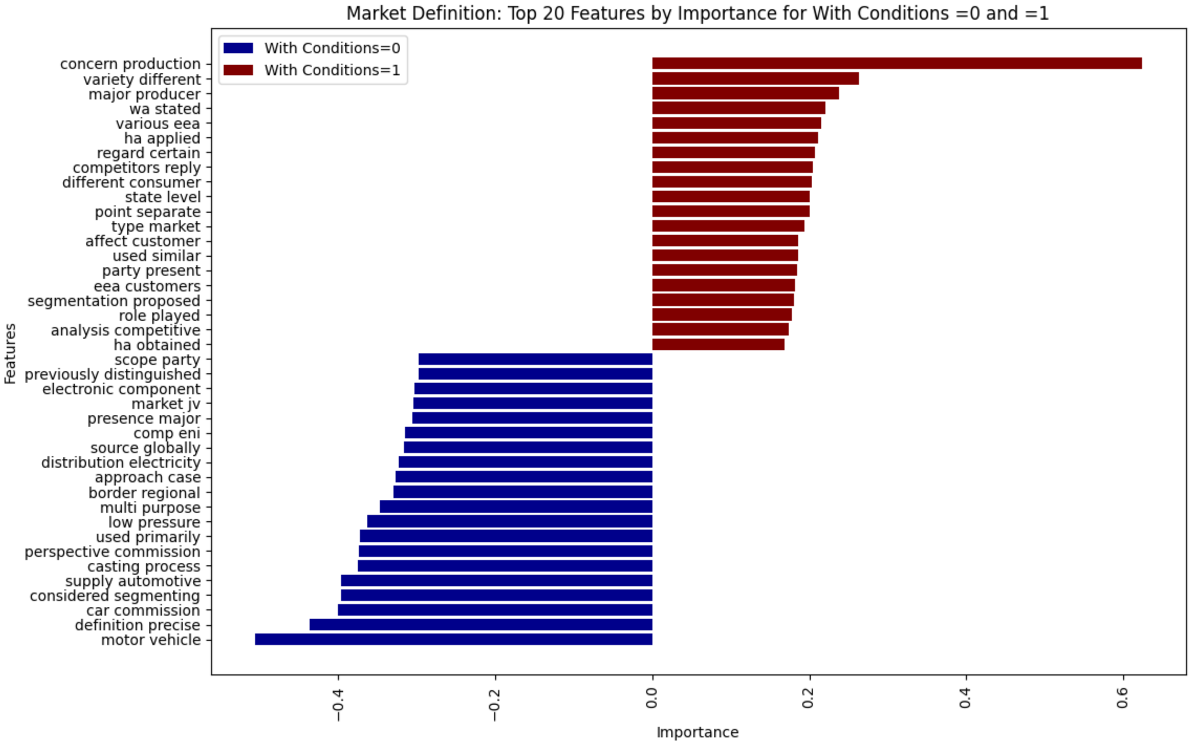


Figure D.3. Top 20 features by importance of *svm_wc_md* per class

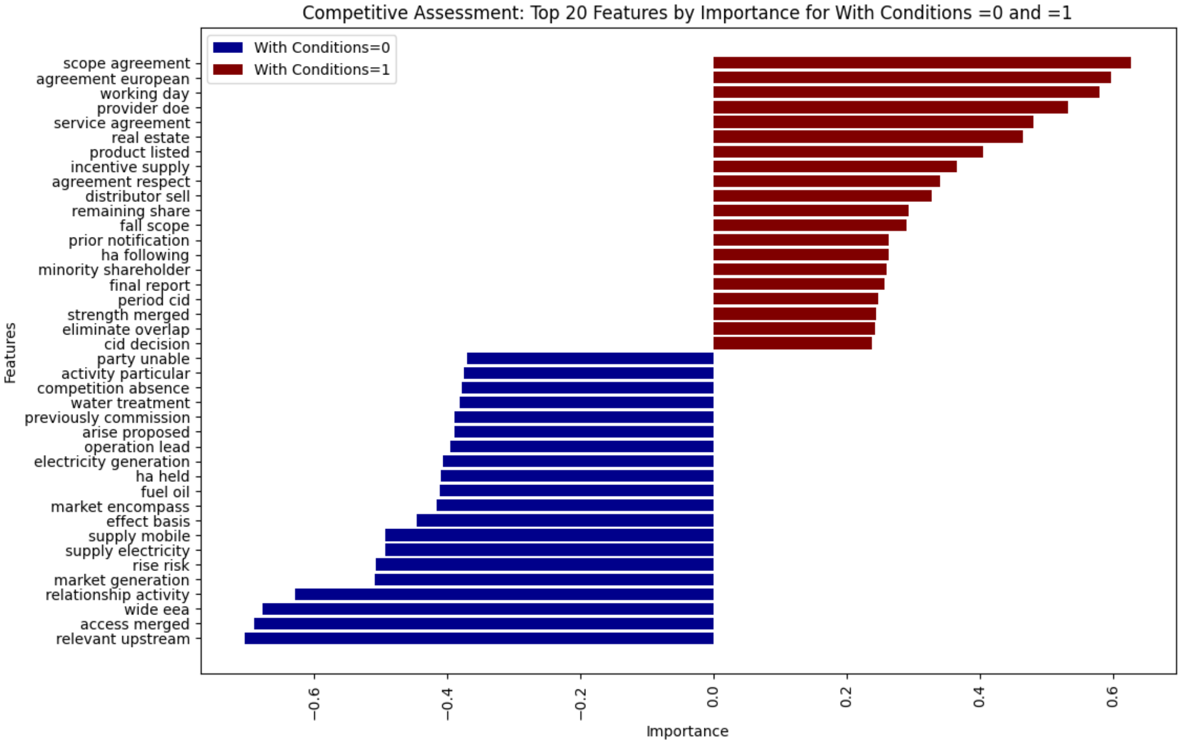


Figure D.4. Top 20 features by importance of *svm_wc_ca* per class

Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on May 3, 2023 is identical to the printed version I submitted to the Examination Office on May 3, 2023.

DATE: May 3, 2023

NAME: Ma. Adelle Gia Toledo Arbo