Active Learning for Financial Investment Reports

Sian Gooding and Ted Briscoe
Dept of Computer Science and Technology
University of Cambridge
{shg36|ejb}@cam.ac.uk

Abstract

Investment reports contain qualitative information from numerous sources. Due to the huge volume of online business information, it is increasingly difficult for financial analysts to track and read all relevant texts. In this paper, we develop a novel tool to assist financial analysts when writing an investment report. We perform multi-class classification on business texts to categorise them into informative investment topics. Using active learning we show that we can obtain the same F1-score of 0.74 with 58% less data.

1 Introduction

Financial analysts guide investors and asset managers in their investment choices (Knorr Cetina and Preda, 2012) by providing investment research information, recommendations, advice or market decisions (Bauman and Dowen, 1988). Such information is typically presented in report format and used by investors to inform portfolio decisions (Baker and Haslem, 1973).

Investment reports contain information from numerous sources and aim to present facts in a coherent and readily intelligible manner (Graham et al., 1934). As well as quantitative measures, investment reports cover a wide range of qualitative topics such as customer satisfaction, brand recognition, and corporate social responsibility (Huang et al., 2014).

Due to the rise of online resources, the availability and accessibility of business information has rapidly increased (Fogarty and Rogers, 2005). Owing to this, it is often infeasible for a financial analyst to keep track of, let alone read, all available information on a given company (Seo et al., 2004).

In this paper we present an automated pipeline to identify and categorise pertinent investment information. We incorporate our models into an active learning framework, allowing financial analysts to train the system with a minimal number of annotated examples. We envision our system being used to assist financial analysts in acquiring and categorising relevant company information.

2 Background

2.1 Financial Text Mining

Prior work on textual classification in the investment domain has extensively focused on the prediction of financial markets (Nassirtoussi et al., 2014). More specifically, algorithms are trained to predict stock price movements using text information from a range of online sources, e.g., the Financial Times, Reuters, or the Wall Street Journal (Cho et al., 1999).

A review by Mittermayer and Knolmayer (2006) compares eight text mining prototypes used for predicting short-term market trends. All prototypes rely exclusively on text-based features. The systems opted for either expertly handcrafted features or features automatically inferred by models. Most of the financial performances obtained by the systems are moderate; Mittermayer and Knolmayer (2006) argue that this is due, in part, to the systems not considering quantitative information. However, they acknowledge that qualitative information is highly informative. For example, when a company reports that it received a 'takeover bid' the crucial data is not in a numerical format.

A further application of financial text mining, similar to the production of investment reports, is that of automated portfolio management. Portfolio management involves the monitoring of current investments by finding, filtering and evaluating relevant information. Warren is a multi-agent system for intelligent portfolio management by Seo

et al. (2004). This system enables users to keep track of both quantitative (e.g., stock price, performance history) and qualitative information in the form of online financial news reports. The text mining component of Warren, referred to as TextMiner (Seo et al., 2002), performs text classification on financial articles. TextMiner uses a combination of word feature sets and a variant of the weighted majority algorithm to classify news articles. Articles are classified into one of five classes, each class aims to represent the financial performance of the company based on the article, for instance *good*, *good-uncertain*, *neutral*. TextMiner achieves a 75% average accuracy across all classes. One difficulty the authors note is that the system struggles when presented with phrases from multiple classes, for example 'Company B shares rose 5% contrasting with A where shares fell by 7%'. Warren uses sets of words as features e.g., 'shares rose','shares fell', but is unable to link these to relevant entities.

Unlike the previous systems presented in Mittermayer and Knolmayer (2006), we do not aim to predict the impact of relevant business information directly on stock prices. Neither do we attempt to classify text according to the financial impact like the Warren system. Instead, our system is designed to present useful and targeted information from a financial analyst's perspective. To the best of our knowledge this is the first system designed for this task.

2.2 Active Learning

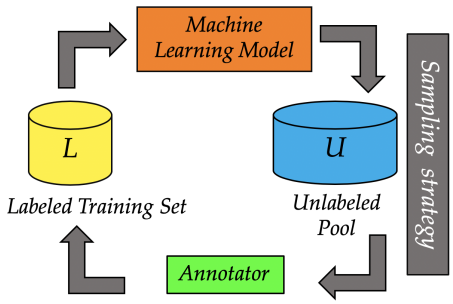


Figure 1: Active learning cycle

Annotated data is hard and expensive to obtain, notably in specialised domains where only experts can provide reliable labels (Konyushkova et al., 2017). Active learning allows machine learning classifiers to achieve higher accuracies with fewer training instances by enabling the classifier to interactively query data points. Active learning is well-motivated in many modern machine learning problems where data may be abundant but labels are scarce or expensive to acquire (Settles, 2009).

Figure 1 shows a classic active learning scenario; whereby a machine learning model has access to an unlabelled pool of data and an *uncertainty sampling strategy* is used to select the most informative instances for labelling. Once the most informative instances have been labelled they are added to the training set and the model is then retrained.

Our motivation for incorporating an active learning framework into the system is two-fold:

1. **Annotator Resource**
   Gathering labelled data for this task is time-consuming and requires the expertise of experienced financial analysts. Maximising the utility of the labelled data allows for better models with fewer labelled instances, saving valuable resources.

2. **Category Introduction**
   When writing financial reports the relevant qualitative categories are subject to change over time. Since new labels may be introduced by financial analysts it is important that the model is able to prioritise acquiring labels for new topics.

3 Data

The data set used in this project was collected by All Street Research[1] ("All Street"), who specialise in creating intelligent tools for financial analysis. It was created using online business resources annotated by financial analysts. Analysts were asked to select information that they would consider useful when writing an investment report. This selected text was then labelled according to the category of the investment report it was relevant to. An example of annotated text from the data set is shown in Table 1.

The total data set collected contained 3097 instances, with individual categories defined by analysts. However, several categories contained less than 100 examples which meant they were not large enough to train and test our framework. We therefore limit the data set to topics that have at least 100 instances. The resulting data set consists of 1824 examples and 11 categories; a breakdown of the categories is shown in Table 2. The category

[1] https://www.allstreet.org

Source: Pfizer 2016 Annual Review

HOSuN fuses our global physical supply chain with a global information supply chain, enabling complete visibility into the status of products at all times. This makes our management of the supply process more efficient. Through HOSuN, we can also use predictive analytics to anticipate future demand patterns. This knowledge is crucial for the efficient production and cost reduction of biologic and vaccine products.

Artificial Intelligence | Cost Reduction | Supply Chain | Not Labelled

Table 1: Example of analyst annotated text

with most examples (340) was *Artificial Intelligence*, with samples of text covering many areas such as 'data mining', 'machine learning' and 'big data'. The smallest category was *Wellbeing* consisting of 196 examples. The mean word length across examples in each topic is reported; the category *Human Capital* had the highest average word count (575) and *Culture* the lowest (380).

Category	Total	Mean Length
Artificial Intelligence	340	430
Business Process Innovation	137	426
Climate Action	228	557
Cost Reduction	120	416
Culture	106	380
Customer Service	160	555
Enterprise Solutions	129	425
Human Capital	119	575
Quality Education	109	532
Supply Chain Management	180	393
Wellbeing	196	476

Table 2: Data set categories alongside the total number of examples and the mean word length

4 Method

Our classification pipeline consists of three steps, which are embedded into an active learning framework. The classification pipeline is outlined in Section 4.1, and the active learning settings are described in Section 4.2.

4.1 Topic Classification Pipeline

4.1.1 Preprocessing

The first stage of classification involves preprocessing the text. In the samples provided we initially remove any corporate named entities, names of people and stop words using spaCy.[2] In the wild, our system is provided with the URLs of relevant web pages; text is then scraped from the page and the pre-processing is performed on paragraph content. Irrelevant content such as page headings are disregarded at this stage.

4.1.2 Feature Selection

Our system relies on word features as it aims to identify terms or bigrams that are highly indicative of a given class. We use functions from the scikit-learn[3] library to transform the total vocabulary of our training set to a matrix of token counts. We then apply a scikit-learn transformer in order to produce a normalized *tf-idf* representation of content. This technique is a common term weighting scheme used in information retrieval and document classification. The goal of using *tf-idf* instead of raw word frequencies is to minimise the impact of highly frequent tokens across a corpus, thereby maximising the importance of class-discriminative terms. Using this technique we are able to investigate which terms are most discriminative for a given class. Examples of the most informative terms for the *Artificial Intelligence* and *Climate Action* classes are shown in Figure 2.

4.1.3 Model Selection

We tested a range of multi-class models using stratified 5-fold cross-validation. The average macro F1-score across all classes is reported for the top three performing classifiers in Table 3.

Classifier	F1-score
Linear SVC (calibrated)	0.74
Linear SVC	0.72
Logistic Regression	0.71
Random Forest	0.69

Table 3: Results

The best performance on this data set was by the linear support vector (SVC) model. Cali-

[2] https://github.com/explosion/spaCy

[3] https://scikit-learn.org

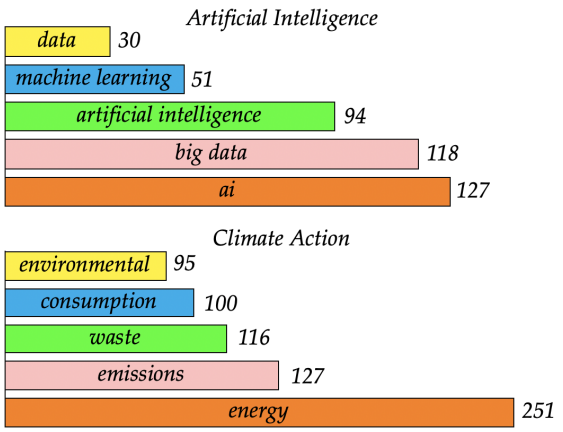


Figure 2: Terms with highest *tf-idf* value for the classes *Artificial Intelligence* and *Climate Action*, shown with class occurrence counts

brated SVC results are obtained using a cross-validation estimator which enables automatic hyper-parameter selection using cross-validation on the training set. The best parameter settings across 5 folds are averaged for prediction on the test set. A more in-depth analysis of classifier results is presented in Section 5.

4.2 Active Learning

As outlined in Section 2.2, uncertainty based active learning requires an *uncertainty sampling strategy* (Lewis and Gale, 1994). This strategy allows an active learner to query the instances that it is least certain about labelling (Settles, 2009). We use three uncertainty sampling strategies, described below, and compare their effectiveness. In the following, x^* denotes the most informative instance from an unlabelled set. To illustrate the sampling strategies we reference a three class example with two data points, shown in Table 4.

Data	Class 1	Class 2	Class 3
1	0.60	0.40	0.00
2	0.50	0.25	0.25

Table 4: Example multi-class probability distribution for two data points

4.2.1 Least Confidence Sampling

This technique considers which of the unlabelled instances has the lowest maximum confidence (Lewis and Gale, 1994):

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\, 1 - P_\theta(\hat{y}|x),$$

where $\hat{y} = \operatorname{argmax}_x P_\theta(y|x)$, or the class label with the highest posterior probability under the model θ.

For instance, of the two data points in Table 4 the highest probability across classes is 0.60 and 0.50 for 1 and 2 respectively. Data point 2 has the lowest maximum confidence and therefore the active learner would request this label.

4.2.2 Margin Sampling

Multi-class margin sampling (Scheffer et al., 2001) considers the two highest class probabilities $\hat{y}_1$ and $\hat{y}_2$:

$$x^*_M = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x),$$

If there is a large margin between $\hat{y}_1$ and $\hat{y}_2$ then the model is able to discriminate clearly. However, if there is a close margin the model is unsure which class to choose making x a good candidate for labelling.

In our example, the highest two probabilities for point 1 and 2 are $0.60, 0.40$ and $0.50, 0.25$. The difference between these is lower for point 1, therefore the label for this instance should be queried.

4.2.3 Entropy Sampling

The final sampling technique considered uses *entropy* (Shannon, 1948) as an uncertainty measure:

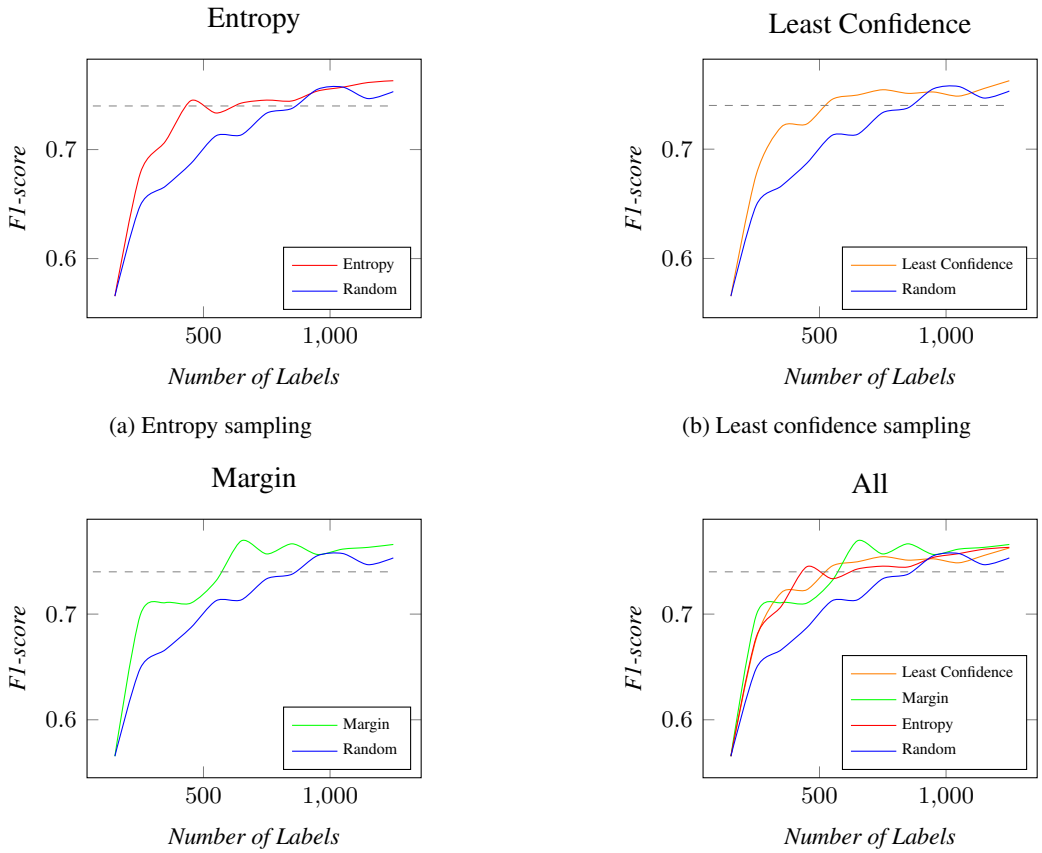
$$x^*_H = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x),$$

where y_i ranges over all possible labels. *Entropy* is an information-theoretic measure that numerically represents the amount of information needed to "encode" a distribution. Entropy is commonly used as an indication of uncertainty or impurity in machine learning (Settles, 2009). For the example in Table 4, the entropy value for point 1 is 0.67 whilst the value for 2 is 1.04. Therefore, point 2 having the highest entropy value would be chosen for labelling.

5 Results

5.1 Active Learning Results

In this section we present the results for each *uncertainty sampling strategy*. To compare the impact of intelligently selecting data for labelling, these techniques are presented alongside a random baseline. The baseline represents the average performance across 5 runs with random data



(a) Entropy sampling

(b) Least confidence sampling

(c) Margin sampling

(d) All sampling approaches

Figure 3: Active learning results

sampling. The classification model used in all settings is the *calibrated SVC*, as this was the best performing model shown in Section 4.1.3. In order to test how effective the active learning techniques would be in practice, we simulate annotation by withholding labels from our current data set and provide them when the active learner queries for the label. The number of labels provided is shown along the x axis. The initial model is trained with 150 random labelled instances; the model is then retrained with additional labels requested by the uncertainty sampling strategy. Once retrained, the F1-score is calculated using a held-out test set of size 548.

Entropy, as shown in Figure 3a, is a highly successful uncertainty sampling approach. The dashed line marks an F1-score of 0.74, as this was the best score achieved with 5-fold cross-validation on the total data set. Using entropy sampling the model is able to achieve an F1-score of 0.74 with only 448 labelled examples. As the initial model is trained with a random 150 instances, only 298 labels are requested by the classifier to reach this score. In comparison, the random baseline requires 710 additional data points. This means our active learner can achieve the same score with 42% of the labelled data needed by a non-active classifier.

Least confidence sampling, illustrated in Figure 3b, achieves an F1-score of 0.74 with only 313 additional labels. As for entropy-based sampling, the initial improvement gradient is steep. Within the first 200 additional labels, the model improvement using least confidence sampling is 0.18, which is double the improvement of the baseline 0.09.

As shown in Figure 3c, margin sampling achieves an F1-score of 0.74 with 382 additional labels, the most labels required of all active techniques for this score. However, the initial improvement gradient is the highest of all sampling strategies. Furthermore, margin sampling reaches an impressive F1-score of 0.77 with 486 labelled items, surpassing the results of all other techniques and the baseline.

Figure 3d shows all three uncertainty sampling approaches and the random baseline. The sam-

Category	Precision	Recall	F1-score
Artificial Intelligence (104)	0.79	0.84	0.81
Business Innovation (41)	0.62	0.53	0.57
Climate Action (70)	0.96	0.93	0.94
Cost Reduction (35)	0.70	0.74	0.72
Culture (26)	0.72	0.69	0.71
Customer Service (51)	0.90	0.88	0.89
Enterprise Solutions (38)	0.62	0.53	0.57
Human Capital (35)	0.81	0.86	0.83
Quality Education (28)	0.77	0.86	0.81
Supply Chain (58)	0.73	0.79	0.76
Wellbeing (62)	0.88	0.79	0.83

Table 5: Precision, recall and weighted F1-score across classes in the test set

pling strategy that reached an F1-score of 0.74 first was entropy-based, followed by least confidence and then margin. All techniques exhibit a degree of variance during retraining, resulting in performance peaks and troughs. To counteract this, our framework monitors performance and saves the best performing models.

5.2 Model Results

The highest F1-score of 0.77 is achieved using margin uncertainty sampling with 747 labelled instances. Comparatively, the highest baseline score is 0.76 and requires 1216 labelled instances. The reason the random baseline does not achieve an F1-score of 0.77, even when trained with the total data set, may be due to the fact that the calibrated SVC re-tunes optimal parameters at each training step. Therefore, parameters for all models will depend on the order of labels they were presented with.

Table 5 presents the performance across classes for this model. A confusion matrix is provided in Appendix A.1. The best performance is achieved on the *climate action* class where 65 of the 70 instances in the test set are labelled correctly. The worst performance is on *Business Innovation* and *Enterprise Solutions*, both with a weighted F1-score of 0.57. A closer inspection of the misclassifications for these classes provides an insight into why performance declines. For instance, consider example (1):

(1) We fuse our global supply chain with an information supply chain, enabling complete visibility into the status of products at all times. In turn making our management of the supply process more efficient.

This has been attributed the label *enterprise solution* and is misclassified into the *supply chain* category. This raises the question of whether segments of text could be attributed multiple labels in future labelling scenarios if they are relevant to multiple classes.

6 Conclusion

To conclude, we have built a classification pipeline that can be used with online business resources to categorise investment-related content. The pipeline is incorporated into an active learning framework, allowing financial analysts to train effective models with as few as 448 labelled examples. Our best performing active learning model achieves an F1-score of 0.77 with 747 instances. In practice there would be a much larger unlabelled data set, allowing the model more variety and choice when requesting data to be labelled.

In future work we aim to integrate additional features into our topic classification pipeline, as well as test our active learning loop in the wild with financial analysts. Further to this, we recognise a drawback of our current approach is that we do not initially filter for content relevancy. Therefore we plan to investigate techniques of disregarding repeated or irrelevant information prior to multi-class classification.

Acknowledgements

We would like to thank All Street[4] and Innovate UK[5] for funding this project and providing the data.

[4] https://www.allstreet.org

[5] https://www.gov.uk/government/organisations/innovate-uk/about

References

H Kent Baker and John A Haslem. 1973. Information needs of individual investors. *Journal of accountancy*, pages 64–69.

W Scott Bauman and Richard Dowen. 1988. Growth projections and commin stock returns. *Financial Analysts Journal*, 44(4):79.

V Cho, B Wüthrich, and J Zhang. 1999. Text processing for classification. *Journal of Computational Intelligence in Finance*, 7(2):6–22.

Timothy J Fogarty and Rodney K Rogers. 2005. Financial analysts' reports: an extended institutional theory evaluation. *Accounting, Organizations and Society*, 30(4):331–356.

Benjamin Graham, David Le Fevre Dodd, Sidney Cottle, et al. 1934. *Security analysis*. McGraw-Hill New York.

Allen H Huang, Amy Y Zang, and Rong Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review*, 89(6):2151–2180.

Karin Knorr Cetina and Alex Preda. 2012. *The Oxford handbook of the sociology of finance*. Oxford University Press.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR94*, pages 3–12. Springer.

Marc-André Mittermayer and Gerhard Knolmayer. 2006. *Text mining systems for market response to news: A survey*. Institut für Wirtschaftsinformatik der Universität Bern.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer.

Young-Woo Seo, Joseph Giampapa, and Katia Sycara. 2002. Text classification for intelligent portfolio management. Technical report, Carnegie-Mellon University Pittsburgh PA Robotics Institute.

Young-Woo Seo, Joseph Giampapa, and Katia Sycara. 2004. Financial news analysis for intelligent portfolio management. Technical report, Carnegie-Mellon University Pittsburgh PA Robotics Institute.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Claude E Shannon. 1948. A note on the concept of entropy. *Bell System Tech. J*, 27(3):379–423.

A Appendix

A.1 Model Confusion Matrix

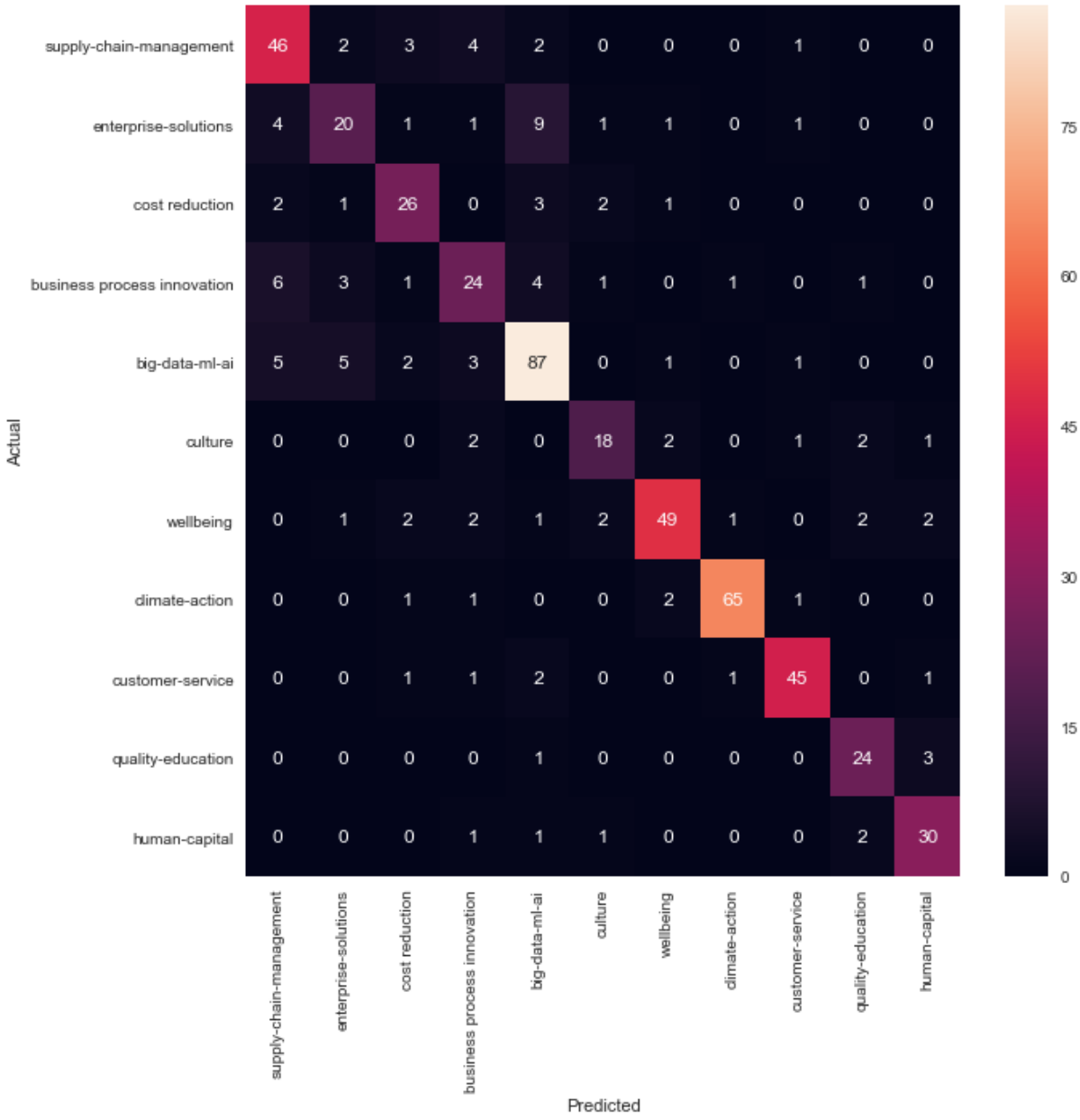


Figure 4: Confusion matrix for best performing margin sampling model