Supplemental Material for Numerical Claim Detection in Finance: A Weak-Supervision Approach

APPENDIX A
EXPERIMENTS OVER MULTIPLE SEEDS

The test accuracy of weak-supervision model and all three variants of BERT for five different seeds are listed in Table 1.

TABLE 1
Accuracy analysis of our model and three BERT models

Seed	WS	BERT-G	BERT-W	BERT-WG
42	0.9404	0.9442	0.9368	0.9442
149	0.9479	0.9591	0.9480	0.9554
1729	0.8996	0.9294	0.8959	0.8922
13832	0.9553	0.9628	0.9480	0.9480
110656	0.9330	0.9740	0.9405	0.9405
Avg.	0.9353	0.9539	0.9338	0.9360

APPENDIX B
FLOWCHART OF OUR METHODOLOGY

Figure 1 gives an overview of the steps involved in the complete pipeline. There are two main steps through which the raw data is passed in order to generate enriched dataset for input to our weak-supervision model. The labelled datasets generated from weak-supervision model and manual annotation are then comprehensively analysed.

APPENDIX C
LABELLING FUNCTIONS METHODOLOGY

The following illustrates the methodology adopted by us while choosing the rules to define the weak-supervision mode. All rules were acknowledged post detailed analysis of sample documents distributed over sector and time :

1) Phrases often provided definitive information about a given sentence in a document and in most cases they had a fairly consistent linguistic composition. It was exploited to both identify out-of-claim and in-claim sentences.
2) Certain phrases such as ”reasons to buy”, ”reasons to sell” or the presence of words which are indicative of past tense such as ”was”, ”were” are characteristic of out-of-claim sentences, since they indicated either facts or events which happened in the past.
3) The alternate adoption of phrase matching was to identify in-claim sentences. This mostly consisted of a verb form indicative of a probabilistic event (eg: likely, intends) coupled with preposition (usually ”to” or ”at”). Based on the ambiguity of the resulting phrase they were either categorised as a high-confidence claim or a low-confidence one.
4) In a bid to capture the effect of a few other verb forms indicative of a probabilistic event, we also chose to look at its lemmatized form to reduce inflectional usage and use the base token for a more holistic evaluation over multiple usage formats.
5) POS tags were also derived for ”project” as a word wherever present. This was done to segregate its usage as a verb. Its usage as a verb was usually observed to be adopted while making claims or predictions.

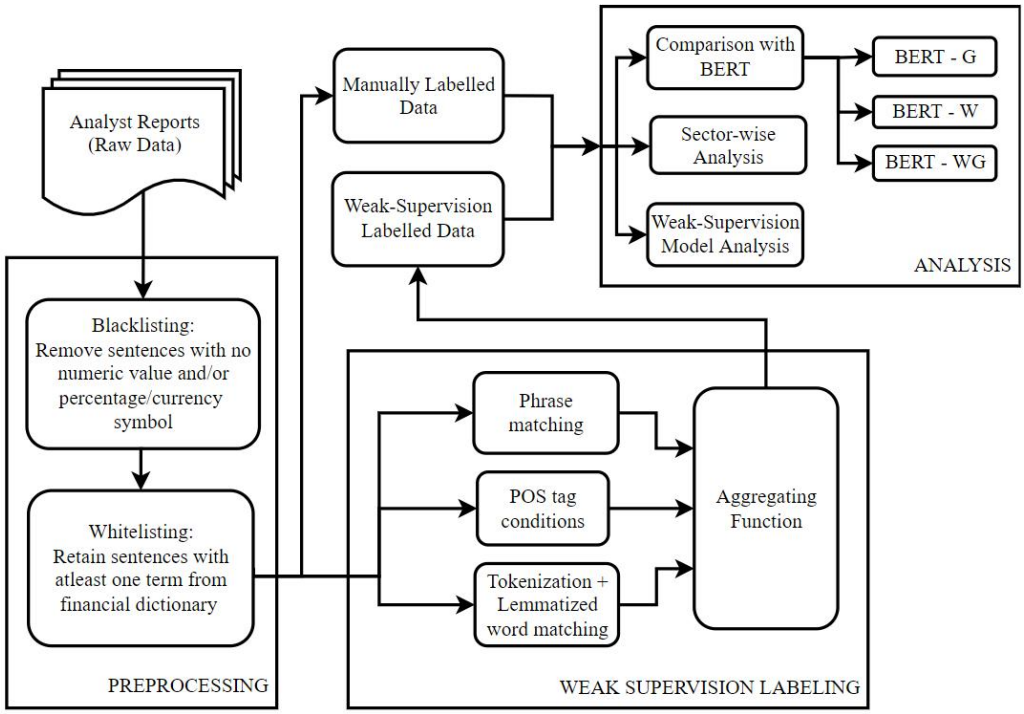

Fig. 1. Flowchart for complete methodology