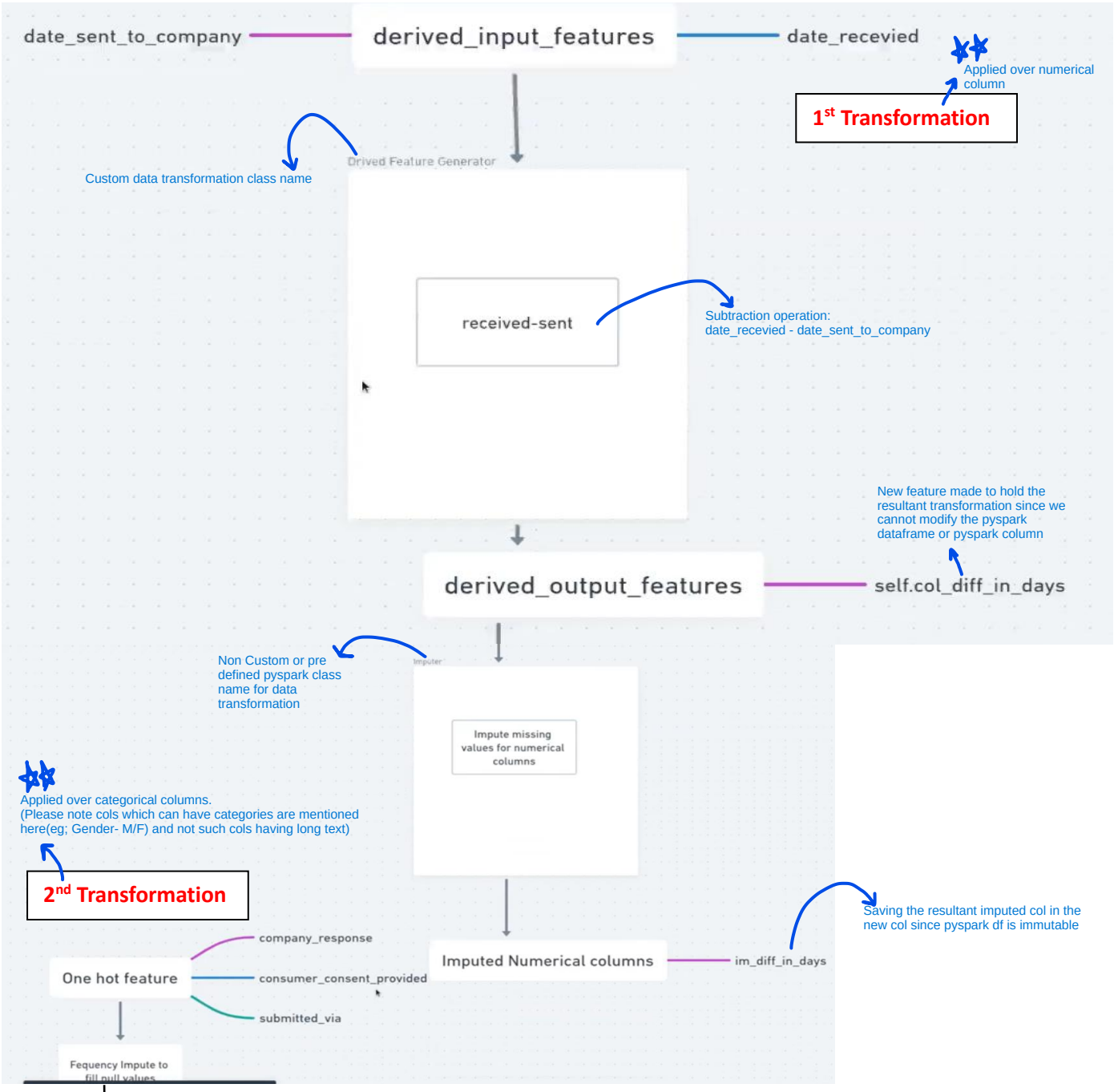
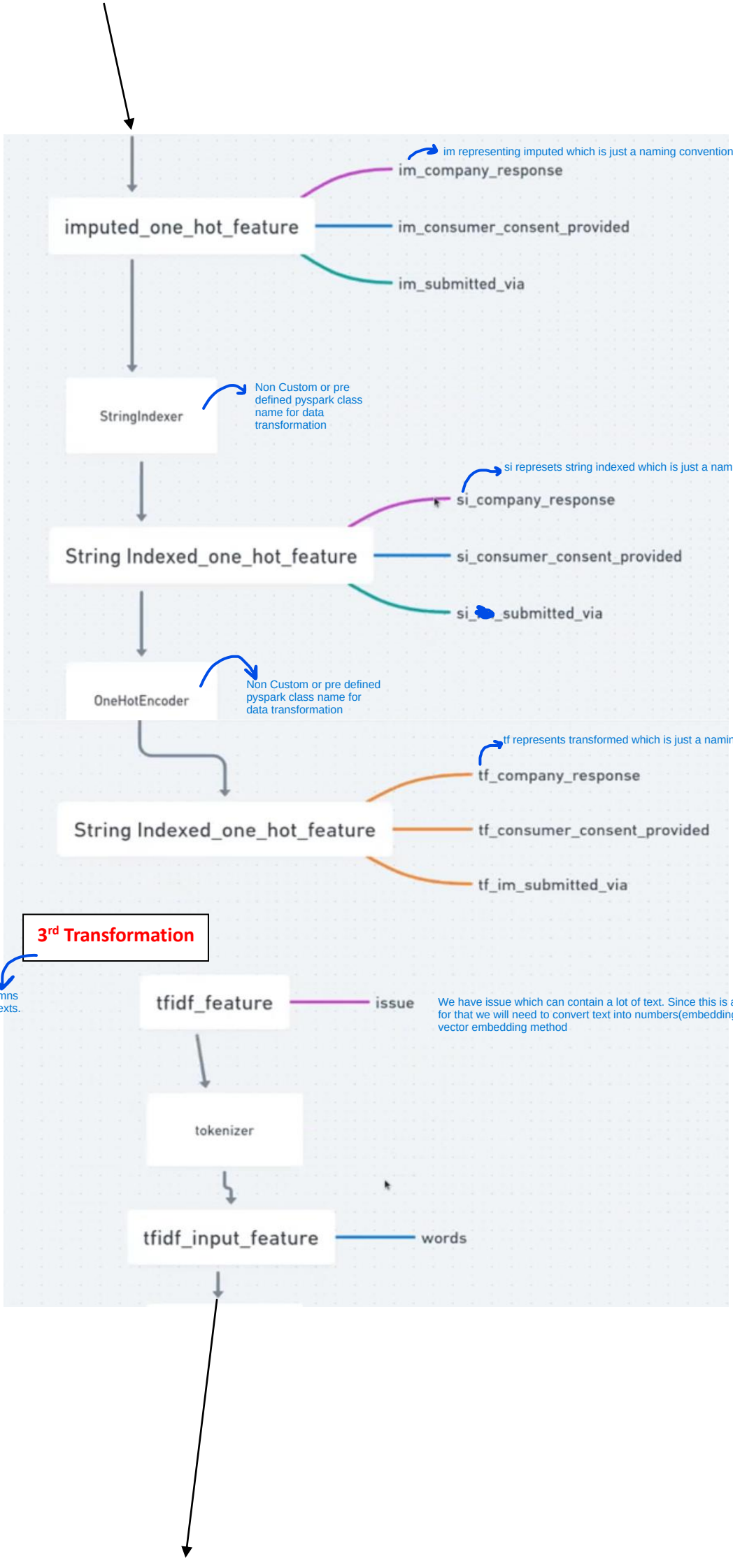
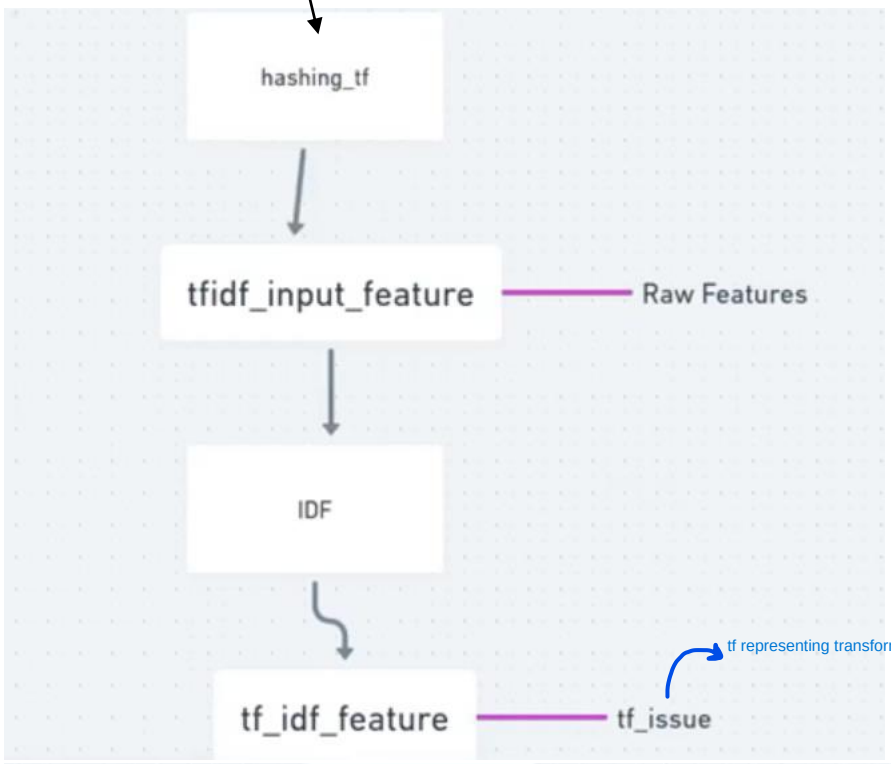
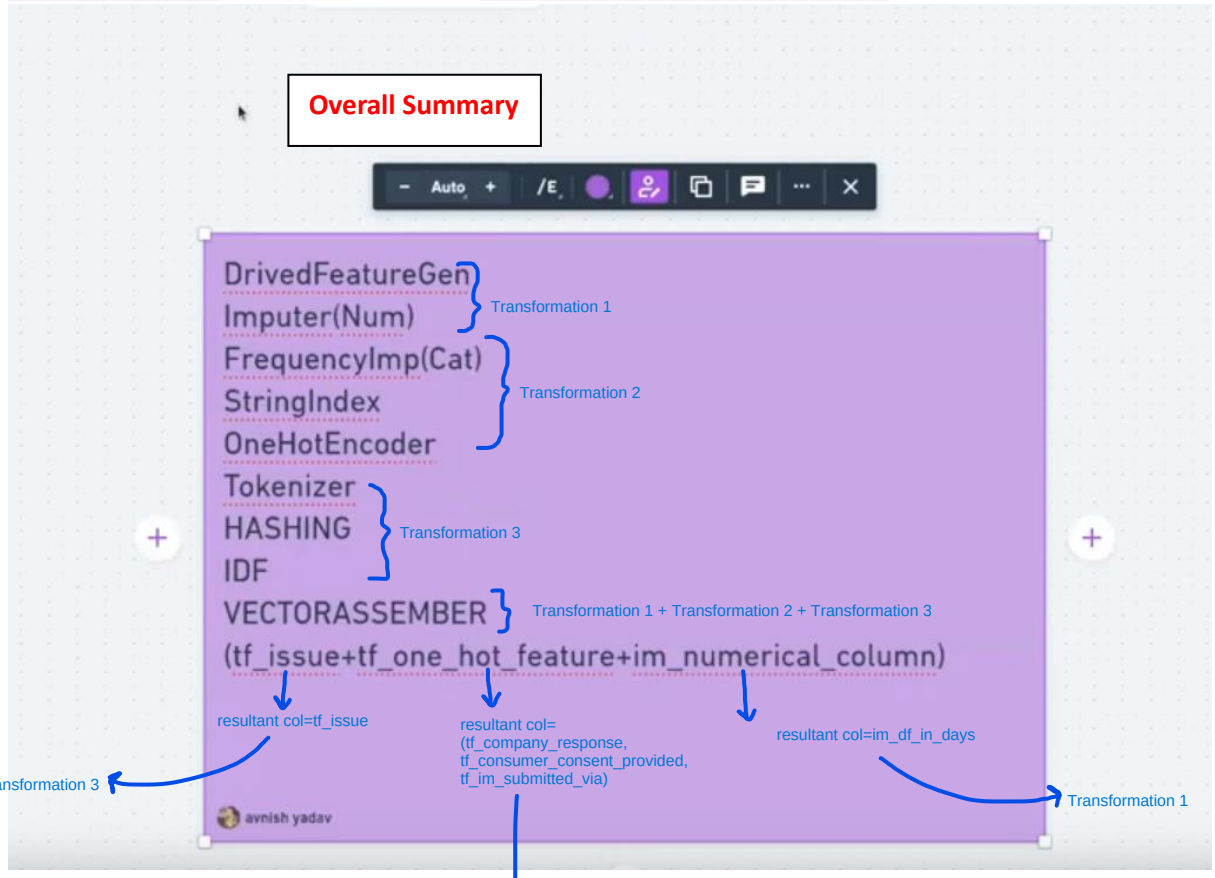
::::::::::******Arranging different transformations done using Pipeline flowchart*******::::::::::




im representing imputed which is just a naming convention defined to store the transformed col in new col
im_company_response
imputed_one_hot_feature
im_consumer_consent_provided
im_submitted_via
StringIndexer
Non Custom or pre defined pyspark class name for data transformation
si represets string indexed which is just a naming convention
si_company_response
String Indexed_one_hot_feature
si_consumer_consent_provided
si_submitted_via
OneHotEncoder
Non Custom or pre defined pyspark class name for data transformation
tf represents transformed which is just a naming convention
tf_company_response
String Indexed_one_hot_feature
tf_consumer_consent_provided
tf_im_submitted_via
3rd Transformation
Applied on columns that have long texts.
tfidf_feature
issue
We have issue which can contain a lot of text. Since this is an important feature we need to include it. And for that we will need to convert text into numbers(embeddings) which we are achieving through tfidf text to vector embedding method
tokenizer
tfidf_input_feature
words


hashing_tf
tfidf_input_feature
Raw Features
IDF
tf_idf_feature
tf_issue
tf representing transformed


Overall Summary
DrivedFeatureGen
Imputer(Num)
Transformation 1
FrequencyImp(Cat)
StringIndex
OneHotEncoder
Transformation 2
Tokenizer
HASHING
IDF
Transformation 3
VECTORASSEMBER
Transformation 1 + Transformation 2 + Transformation 3
(tf_issue+tf_one_hot_feature+im_numerical_column)
resultant col=tf_issue
Transformation 3
resultant col=
(tf_company_response,
tf_consumer_consent_provided,
tf_im_submitted_via)
Transformation 2
resultant col=im_df_in_days
Transformation 1
avnish yadav