Exploring Concept Drift in Technology by Tweets Mining

Alptekin Durmusoglu
Associate Professor, Department of Industrial Engineering
Gaziantep University
Gaziantep, Turkey
durmusoglu@gantep.edu.tr

Mohamad Nasi
M.Sc. Student, Department of Industrial Engineering
Gaziantep University
Gaziantep, Turkey
ma11012@mail2.gantep.edu.tr

Abstract

Over the last decade, a dramatic transform happened in information sources and their use in the digital era. Social media networks have brought a new way of expressing the sentiments of individuals. The matter went beyond being an expression of separate opinions of some individuals, as companies, official institutions and various organizations have pages on the communication sites through which they share various developments, products, opinions, and sometimes even official decisions. Social media become a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as a key factor in decision making. Twitter is a microblogging service built to describe what is happening anywhere worldwide, at any moment. It's a fascinating forum for more than 500 million messages per day from about 1.3 billion people. Twitter data is short, specific, and easily accessible, that's why it has become one of the best sources for sentimental analysis and knowledge discovery by data streams mining. One of the major issues that affect data streams mining is that the underlying distribution of data may change over time leading to the phenomenon of concept drift. Identification and handling concept-drift in Data Streams is present area of interest. In this study, we present an approach to explore and understand the concept drift occurring in Twitter data streams. Two machine learning technique Naive Bayes Classifier and Extreme Gradient Boosting (XGBoost) Classifier were applied on more than 300K tweets from International Technology Companies, and to detect / understand concept drift and specify whether concept drift in a technology area is a radical or an incremental innovation.

Keywords

Tweets Mining, Social Media Analysis, Concept Drift, Data Stream, Technology Tweets.

1. Introduction

The acceleration in the volume of data produced daily is growing constantly due to the increased number of applications that generate big amounts of data at a great velocity, such as the content and reactions on social media, advertisements click, shares, transactions, streaming content, the Internet of Things (IoT) realms and so much more. The need to deal with huge amounts of data has motivated the research in the field of data mining aiming to generate knowledge from data, which can be translated as discovering of new and non-trivial patterns, relations, and trends in data useful to the user (Schuh, et al. 2019)

Over the last decade, a dramatic transform happened in information sources and their use in the digital era. Social media networks have brought a new way of expressing the sentiments of individuals. The matter went beyond being an expression of separate opinions of some individuals as companies, official institutions and various organizations have pages on the communication sites through which they share various developments, products, opinions, and sometimes even official decisions. Social media become a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as

a key factor in decision making.

Twitter is a microblogging service built to describe what is happening anywhere worldwide, at any moment. Twitter data is short, specific and easily accessible, that's why it has become one of the best sources for sentimental analysis and knowledge discovery by data streams mining. Sentiment analysis (SA) or opinion mining (OM) is the study of public opinions, sentiments, attitudes, and emotions expressed in social media (Bhuvaneswari and Srividhya 2017). The sentimental analysis is an intricate process, which consists of several tasks such as sentiment analysis subjectivity analysis, opinion mining and sentiment orientation. It is considered a novel, evolving new research field in machine learning (Saberi and Saad, Sentiment Analysis or Opinion Mining: A Review 2017)

One of the major issues that affect data streams mining is that the underlying distribution of data may change over time leading to the phenomenon of concept drift. Concept Drift Analysis is the integrated study of identifying and handling Concept-Drift in this evolving stream of data (M 2015). In a concept drift context, we can discard the old data and retrain the model using new observations (sudden drift) or combine the old data with the new data to update the model (gradual drift) or maintain the model as unchanged (no drift) (Yang, et al. 2022).

1.1 Objectives

This paper aims at exploring and understanding the concept drift occurring in Twitter data streams of international technology companies over specific period (almost 10 years), by applying two machine learning technique Naive Bayes Classifier & XGBoost Classifier, trying to answer the following two research questions

- **RQ1:** Is concept drift in a technology area a radical or an incremental innovation?
- **RQ2:** Can we understand/explore a technological change (innovation) by analyzing tweets?

The remainder of this paper is structured as follows: First, containing the most important work related to this topic. After that, a section explaining the methodology of this paper will be provided. In the next section, the results of the machine learning algorithm will be presented, followed by a separate conclusion and a results discussion part.

2. Literature Review:

2.1 Sentiments Analysis:

Sentiment Analysis (SA) -or opinion mining (OM)- is a branch of text mining. It is a widely used text classification tool that analyzes the source text and discovers the underlying sentiment (predicting people's feelings or emotions about something) which can be either positive, negative, or neutral. The analysis can be done from different directions such as Natural Language Processing methods, application of lexicons with annotated word polarities, along with some machine learning-based approaches.

2.1.1 Sentiment Analysis Approaches:

There are two widespread groups for sentiment analysis:

A. Machine Learning Approach:

Depends on machine learning techniques to give information about the polarity of sentiments. To perform classification, two collection of text document are needed: training collection and test collection; the first one is used by the classifier to differentiate between text features, and the second one is used to discover the accuracy of classification/prediction. There are many machine learning algorithms used for sentiment analysis, however the performances of Support Vector Machines, Naive Bayes, and Maximum Entropy with SA and classification are highly successful (Saberi and Saad, Sentiment Analysis or Opinion Mining: A Review 2017). Other approaches include K-Nearest Neighbor, Random Forests, and XGBoost. Many researchers have compared these approaches on text data to find the best classifier in sentiment analysis tasks, Agarwal et al. found that Naïve Bayes classifier obtained good results compared to Support Vector Machine on reviews in Cantonese (Agarwal, et al. 2011).

Machine Learning approaches are usually classified as:

1. **Supervised Learning:** requires well-defined and well-labelled corpus as a training set for the model, and another data set to test the model performance & accuracy, many ML algorithms go under this category such as: Support Vector Machine, Random Forests, XGBoost, Naïve Bayes and other classifiers.

2. **Unsupervised Learning:** in this method, no training data set is needed, only input data set is required, and unsupervised methods can be either machine learning based, or lexicon based, an example of these methods is clustering, in which semantic orientation approach is used and algorithms will extract the phrases that include adjectives or adverbs to predict the sentiment orientation.
3. **Semi-Supervised Learning:** this can be considered as a middle solution between the two previous ones, where a series of labelled and unlabeled data is provided, for the purpose of classification.

B. Lexicon-Based Approach:

Depending on an unsupervised learning technique since no training is required under this approach. It can be said that this method determines whether the term is far or close to being positive or negative, that is being done depending on lexical rules. Some researchers introduced a sentimental lexicon (Abbasi, Chen and Salem 2008), the authors argued that that the preparation of manual lexicon is higher effective than the preparation of an automatic sentiment lexicon. Under Lexicon-Based approach there are two categories:

1. **Dictionary-based approach:** in which small set of known-orientation words are collected manually, then this set is increased by searching of well-known corpora for their antonyms and synonyms (Hu and Liu 2004).
2. **Corpus-based approach**: this method depends on syntactic patterns which come together along with seed list of opinion words, aiming to discover other opinion words within big corpus.

C. Hybrid Approach:

lexicon-based machine learning approach that includes manual-written linguistic rules. In this method, cascaded

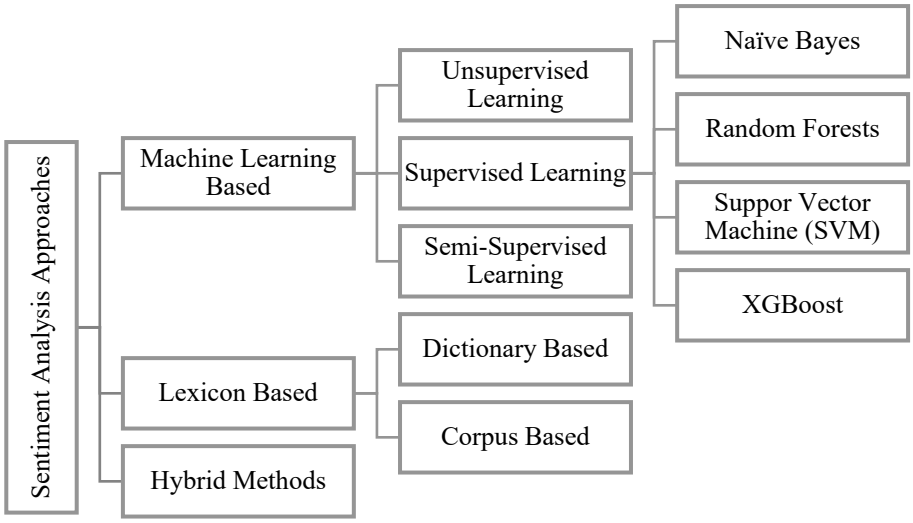

Figure 1 - Sentiment Analysis Approaches

classifiers are used so if one of them failed, the next one performs the classification task, and so on till the categorization of text/document is finished.

2.1.2 Sentiment Analysis of Twitter Data:

Social media is exceptionally useful source which generates huge amount of sentiment rich data including tweets, online blogs, status updates…etc. Twitter is a microblogging service built to describe what is happening anywhere worldwide, at any moment. It is a fascinating forum for more than five hundred million messages per day from about 1.3 billion people. Twitter data is short, specific, and easily accessible, that's why it has become one of the best sources for sentimental analysis and knowledge discovery by data streams mining. (Figure 1)

Sentiment analysis of Twitter data is challenging work due to the use of slang, misspells, abbreviations and the unstructured nature of data which required a lot of preparation and cleaning work on data set to be ready for machine

learning approaches.

The Naive Bayes technique was used in (Gangawane and Torvi 2017) to assess sentiments on abbreviations and short sentences, to improve the algorithm for identifying sentiment words, and to successfully handle bipolar sentiments based on Twitter data.

Many of works related to sentiment analysis have focused on product or movies reviews (Devi, Bai and Ramasub 2020), (Ogul and Ekmekciler 2012), (Wang and Liu 2017), (Dave, Lawrence and Pennock 2003) and (Alencia, et al. 2018) on customer tweets, review sites, blogs or other websites.

(Zhou, et al. 2013) have suggested a Tweets Sentiment Analysis Model (TSAM) that may identify societal interest in and attitudes on the 2010 Australian federal election. The research has shown that it is feasible and advantageous to construct an intelligent system for sentiment analysis based on a lexicon. In (Vasudevan 2017) the sentiment of twitter data was performed depending on Support Vector Machine (SVM) and Decision Tree (DT), and that was performed depending on a dataset consists of 7156 tweets classified with respect to Google self-driving cars, achieving high accuracy with SVM (90%) and less accuracy with DT (65%).

Similar work was introduced in (Barzenji 2021) Whereas sentiment analysis on Twitter text was used to learn about the subjective polarity of the writings, this analysis was carried out using three distinct machine learning algorithms: Support Vector Machine, Radom Forests, and Gaussian Naive Bayes. The classifiers obtained accuracy of 89%, 88%, and 72%, respectively. Similar method was employed in (Gupta, Pruthi and Sahu 2017), where part of that process on sentiment analysis of Twitter data using a mix of two machine learning algorithms, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). According to the findings, the suggested model enhanced the accuracy and f-measure of tweet class prediction.

2.2 Concept Drift Definition & Types:

Concept drift has become a common area of research in the field of data mining and machine learning since it's a major issue when dealing with data streams. Some previous works has focused on the definitions and types of concept drifts, such as:

In (Elwel and Polikar 2011) *concept drift* refers to a change in the class (concept) definitions over time, and therefore a change in the distributions from which the data for these concepts are drawn. The classification of changes in data distribution into two class; real concept drift and virtual concept drift (Widmer and Kubat 1996). The formalization of these two types of drifts as follows (Janardan 2017) and (GAMA, et al. 2014): (Figure 2)

1. *Real Drift*: the posterior probability $P_t(x|y)$ varies over time, independently from variations in the evidence $P_t(x)$.
2. *Virtual Drift*: the evidence or the marginal distribution of the data, $P_t(x)$, changes without affecting the posterior probability of classes $P_t(y|x)$.

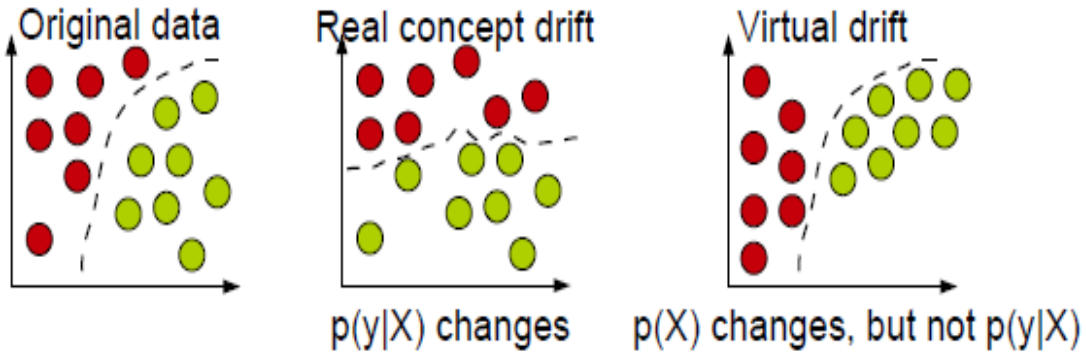

Figure 2- Types of drifts: circles represent instances, different colors represent different classes

Another concept drift is based on how the drift evolves in the system, The drift may occur suddenly, incrementally, or gradually. For better demonstration of the mentioned types, the term "intermediate concept" was introduced by (GAMA, et al. 2014) and to describe the transformation between concepts. Figure 3 illustrates this classification (Lu, et al. 2019):

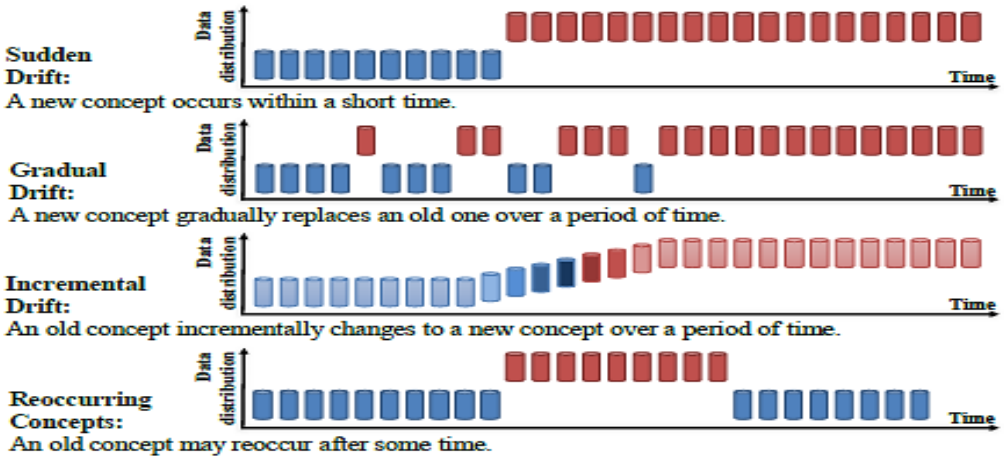

Figure 3- Concept Drift Types

2.3 Concept Drift Detection:

The number of surveys, overviews, and reviews on idea drift detection that were published between 2009 and 2019 was provided by the authors in (Gemaq, et al. 2020)., the result is shown in Figure 4:

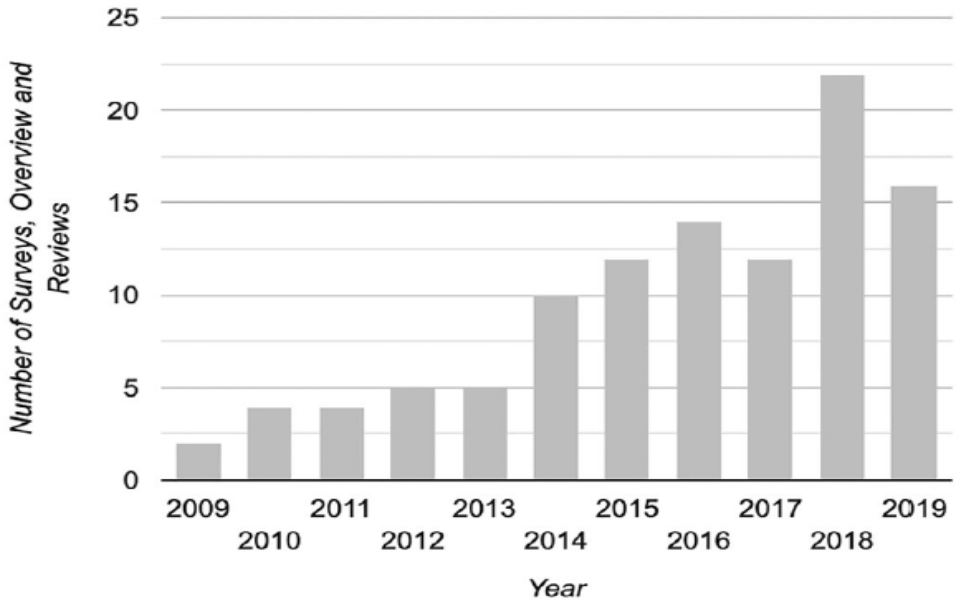

Figure 4- Number of publications on concept drift detectors

Drift detection algorithms can be divided into three categories (Lu, et al. 2019):

2.3.1 Error rate-based drift detection:

The algorithm will monitor changes in the online error rate and will trigger the drift alarm when a substantial change (increase or reduction) of the error is established. Examples of this type of algorithm are:

A. **Drift Detection Method (DDM)** (Gama, et al. 2004), which uses a binomial distribution. That distribution gives the general form of the probability for the random variable that represents the number of errors in a sample of n examples. The concept is modeling the error as a binomial variable, which means that the expected value of the errors can be calculated. DDM monitors the count of errors that resulted from a model learned on the previous stream elements, in general, the model error should stay stable or decrease as more data is used assuming that:
   a. The learning algorithm/method is controlling overfitting.
   b. The distribution of data labels is stationary.

B. **Early Drift Detection Method (EDDM)** (Baena-Garcıa, et al. 2006), which was developed to improve the detection in presence of gradual concept drift and keeps performing well with sudden or abrupt drift detection. The principle behind this method is to focus on the distance between two successive errors' classification instead of focusing only on the number of errors. During the learning process, because the model is developing, the prediction will be improved, which means that the distance between two successive errors will be increasing.

C. **Adaptive Windowing (ADWIN):** (Bifet and Gavaldà 2007): this is one of the popular methods that goes under the window-based detectors, which in general depends on dividing the data streams in a sliding manner, based on either data size or time interval, then the performance of the latest observations will be compared with a reference window. ADWIN uses exponential histograms to detect or estimate the change, it maintains a variable-length window of recent items, ensuring that the data distribution has not changed.

In addition to the methods presented above, there are many other methods that depends on the error rate to detect the drift, such as the one class drifts detection (OCDD) (Gözüaçık and Can 2021) and Page-Hinckley Test (PHT) based drift detector (Qahtan, et al. 2015).

The following Table 1 summarize the error-based classification algorithms (Bayram, Ahmed and Kassler 2022):

Table 1 - Error-based drift detectors

Method	Calculation Method	Tested Hypothesis	Type of drift detected
DDM	Online error rate	Distribution estimation	Sudden drift
EDDM	Online error rate	Distribution estimation	Gradual / Sudden
ADWIM	Error rate difference	Hoeffding bound	Gradual / Sudden
Page-Hinckley	Average value	Performance means	Sudden drift
OCDD	Outlier percentage	Post hoc Neymenvi test	Gradual / Sudden

3.2.2 Data Distribution-based Drift Detection:

Techniques in this category use a metric/distance function to quantify the dissimilarity between historical and current data distributions. Since raw data points are used directly rather than through indirect, abstract information (the learner's output or parameters), data distribution-based drift detection approaches have the advantage of sensitive detection and important output knowledge (when, how, and where concept drift happens).

An example for this kind is: **Competence Model-based drift detection (CM)** (Lu, Guangquan Zhang and Lu 2007) This method is an innovative technique for identifying concept drift in a case-based reasoning system. It presents a new competence model that discovers differences through changes in competence rather than evaluating the actual case distribution. No prior knowledge of case distribution is necessary for the competence-based idea detection method to work, and it offers statistical assurances on the dependability of the changes found as well as accurate descriptions and quantification of these changes.

3.2.3 Multiple Hypothesis Test Drift Detection:

These algorithms apply similar techniques of mentioned in the previous two categories, but they use multiple hypothesis tests to detect concept drift, such as the Linear Four Rate drift detection (LFR) (Wang and Abraham 2015) and Drift Detection Ensemble (DDE) (Maciel, Santos and Barros 2015).

3. Methodology:

This section explains the research steps related to Twitter data collection and processing to make it ready for the next stage where it is analyzed, and machine learning algorithms applied. Twitter was chosen as source of data due to the following reasons: (Figure 5)

- Unlike other social platforms, in Twitter almost user's tweets are completely public, and
- Twitter provides developers and researchers with a streaming API (application programming interface), which allows them to fetch real-time data and past tweets and do complex queries.
- Twitter data is specific, since Twitter API gives the ability to get the tweets related to specific topic or pull the non-retweeted tweets of certain user.

In this study, we've focused on 160K tweets of technology companies over a timeline from 1/January/2015 to 30/June/2019, the data retrieved by Twitter API was pre-processed to remove stop words and punctuations, then stemming/normalization will be applied on tweets to converts the words to their root form (known as lemma). The next stage is sentiment analysis of tweets and splitting dataset to train and test samples with 80/20 ratio so it can be used to train & test the model of Multinomial Naive Bayes Classifier.

Finally, to detect concept drift, a new set of tweets (112K from 1/July/2019 to 25/June/2022) will be fetched to test the Naive Bayes model. Finally, to verify the results, another algorithm known as XGBoost will be applied, and the results will be discussed in the results section.

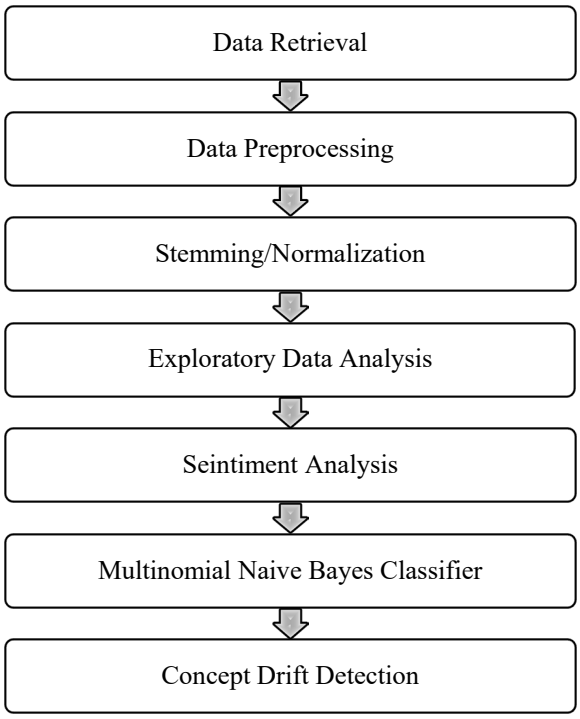

Figure 5 - Work Phases

3. Data Collection:

4.1 Data retrieval

Twitter's Application Programming Interface (API) is a platform that allows researchers and developers to fetch real-time data and historical tweets. API can be used to find and retrieve, engage with, or create a variety of different resources including (tweets, users, trends, media, and others), in addition to the ability to determine the appropriate period. In this study, Twitter API V2 was used to get approximately 300K English tweets which had the following attributes:

- Date
- Username
- Tweet
- Hashtag/Company

The first timeline of tweets fetched is from 1/January/2015 to 25/June/2022.

4.2 Data Pre-Processing

Benefitting from the streaming API tool, the initial Twitter data set has already been filtered to include only English tweets, and retweets – republication or forwarding of an existent tweet – are excluded, the thing that reduces the number of duplicates. The initial number of tweets was 120,015 then became 108,281 after removing the duplications. In the next step, the data was loaded to Jupyter Notebook which is used for Python code of this study. The next step in the pre-processing is removing stop words and punctuation, while keeping the 'emojis' which play an important role in sentiment analysis.

4.3 Stemming/Normalization

Stemming or Normalization is a technique used in Natural Language Processing (NLP) to retrieve information by tracking affixed words back into their root, this will increase the classification accuracy as many studies proved, for example in (Rianto, et al. 2021), the results showed that the accuracy of Support Vector Machine (SVM) classifier has achieved a score of 0.85 when applying stemming compared with 0.73 without stemming. The stemmer used in our study is Porter Stemmer which is the most popular stemmer for English language, and it is known for its speed and simplicity.

5. Results and Discussion

5.1 Sentiment Analysis

The analysis will be performed through the following steps:

A. Detecting Text Polarity:

In this step, the tweets will be labeled with three main values: Positive, Negative and Neutral. These values will be used later to be reflected in the sentiment analysis. The results of this step are shown in Figure 6.

B. Handling Data Imbalance:

It's noticed from the polarity results, that the three categories are imbalanced, and the lowest distribution is about 20%, so the data will be balanced by taking 20% of each category.

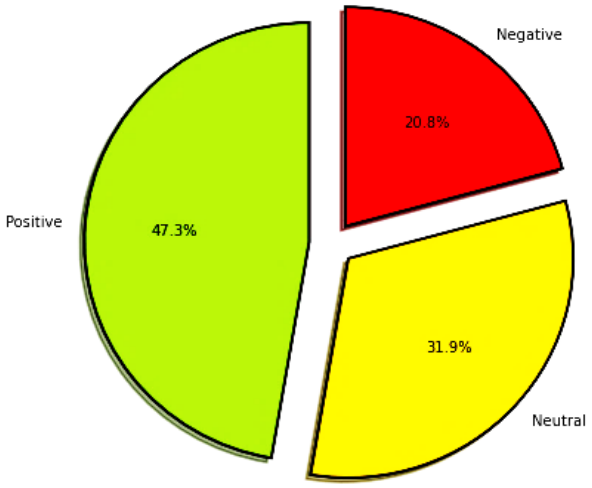

Figure 6 - Tweets Polarities

C. Text Vectorizations:

Machine learning classifiers only deal with numbers, that's why the results obtained from the previous stage should be transformed to a matrix of numbers, this is known as Test Vectorization. There are many ways to perform this step, such as: Bag of Words, (L1) Normalized Term Frequency, (L2) Normalized TF-IDF, Binary Term Frequency and Word2Vec.

In this study, Bag of Words was selected, which is a technique that takes the whole corpus and assigns the vector representation to a given word. The words present are marked as 1 and absent as 0 in the vector representation. The target feature is first encoded using the label encoder. The labels given are: Negative = 0, Neutral = 1, and Positive = 2.

5.2 Multinomial Naive Bayes & XGBoost Classifiers

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Despite its simplicity, Naive Bayesian classifier usually terminates surprisingly well and is widely used because frequently it outperforms more sophisticated classification methods. It's been declared as; Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases (Han, Kamber and Pei 2012).

Before using the Naïve classifier, the data set will be divided into **Train** & **Test** samples by 80/20 ratio respectively. Then the model was built by using the 'Scikit-learn' library in Python. (Figure 7)

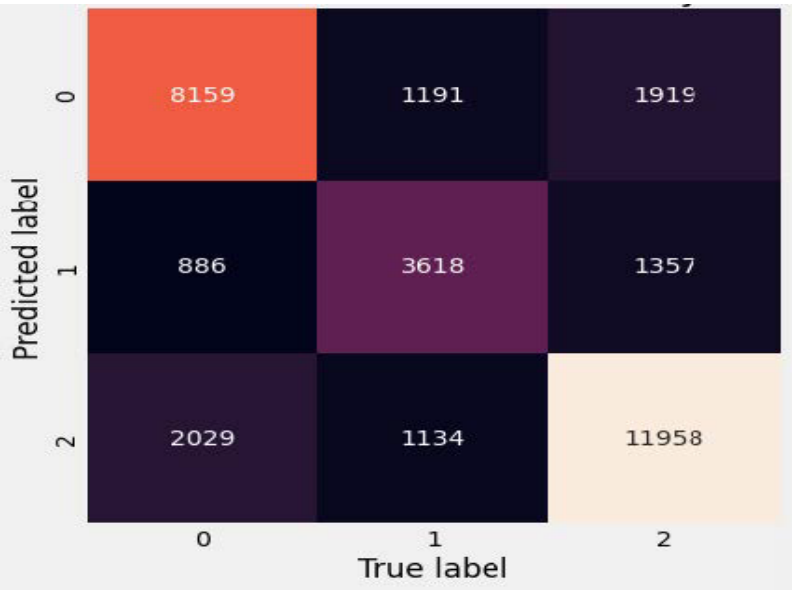

Figure 7- Confusion Matrix for Naive Bayes

According to the confusion matrix, the algorithm correctly predicted 11,952 actual positive tweets (True Positive), 3,618 Neutral tweets (True neutral), and correctly predicted 8,159 negative tweets (True Negative). Given these numbers, the performance metrics were calculated giving the accuracy obtained by the Naive Bayes Model as **73.59%.** (Figure 8)

XGBoost stands for Extreme Gradient Boosting. It's a highly effective machine learning method that depends on gradient tree boosting, which has been shown to give state-of-the-art results on many standard classification benchmarks (Li 2010). XGBoost is a scalable machine learning system for tree boosting the impact of the system has been widely recognized in several machine learning and data mining challenges (Tianqi Chen 2016). In our study, the XGBoost classifier was used, and Figure 8 illustrates the Confusion Matrix of it.
The accuracy obtained by XGBoost Model is **82.14%.**

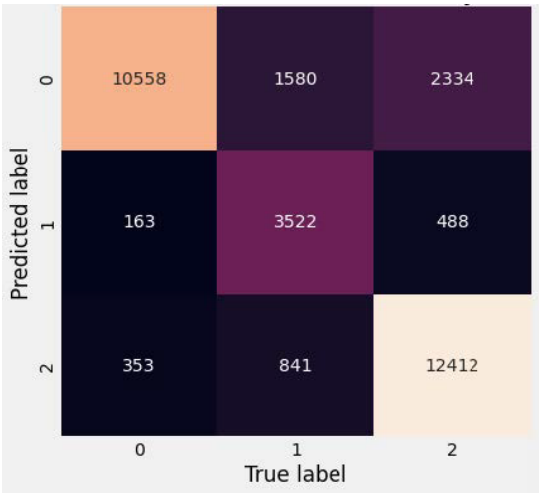

Figure 8- Confusion Matrix for XGBoost

5.3 Concept Drift Detection

This task aims to find out whether the statistical models built on top of twitter data suffer from the problem of label drift or concept drift, to discover it, several steps were followed: (Figure 9)

1. 161,255 tweets were fetched from 01/01/2015 to 29/06/2019, as **training set**. This data was used to train two models – Naive Bayes & XGBoost.
2. 112,411 tweets were fetched from 01/07/2019 to 25/06/2022, as **testing set**. This range of tweets can help determine the concept drift since all the data is sorted by date.
3. In the next step, both the train and test data are concatenated to form a dataset of date wise tweets
4. Finally, a function is defined which splits the data into fragments and check for the accuracy, here, the step size is used as 10000, which means the accuracy is checked for every 10000 samples and the results are stored and further the results are plotted which helps determine the type of drift the model is facing.

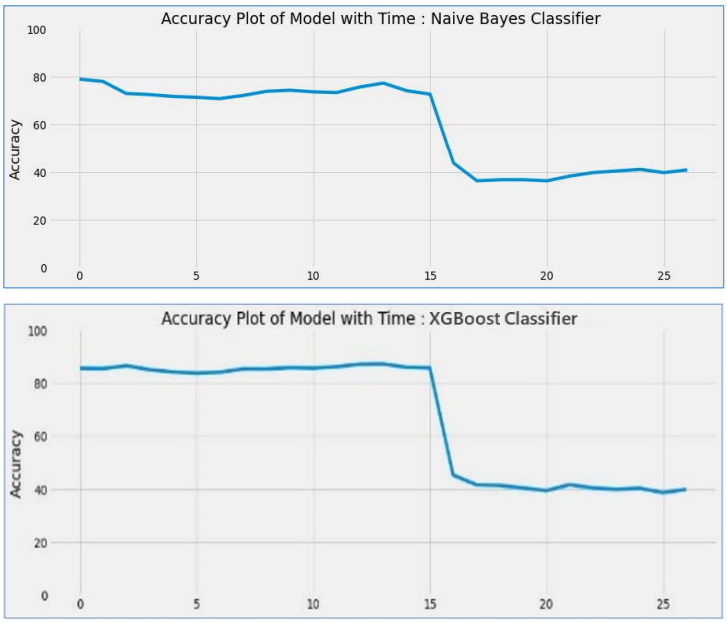

Figure 9 - Accuracy Plot with Concept Drift Effect

Discussion

The model was able to classify the sentiments of tweets into three main classes (positive, negative, and neutral) so the predicted value (the concept) of this model will be distributed over these three classes, with **73.59%** accuracy rate for Naive Bayes and **82.14%** for XGBoost classifiers. However, the monitoring of model accuracy showed that this accuracy ranged from 70-80% on the trained data, but on new data it fell to 30-40 %, which indicates that the model predicted value (the concept) has changed and the result of classification will not be true.

This change in the accuracy rates occurred in a short time and stabled on the new value so, it can be said that the drift the model suffers from is a **sudden drift**, since the accuracy of the first model, Naive Bayes, was suddenly changed from 73% to 30-40%, and to cross verify the existence of sudden concept drift, another classification algorithm (XGBoost Classifier) was implemented and the results were observed to be quite similar. The performance on the model on training data was ~82 % but it fell to ~40 % when the new data was shown to the model.

The sentiment analysis of tweets of technology companies has a larger **positive** fraction, which indicates that: as long as the analysis results is giving the higher positive percentage, the model is not suffering from any drift, on the other hand, receiving different results during the analysis of new tweets will indicates that the predicted value (the concept) is drifted, and there is a change in the topic, product or company that these tweets is focusing on.

The results show that concept drift monitoring can indicates the change in technology by analyzing the tweets of technology companies, however, further analysis needs to be done on the new status (data with drift) to specify whether this change will be considered innovative or not, in other words, whether the new product, idea, solution that led to change, will be accepted by the targeted audience or not. This can be known through new analysis of customers and/or reviewers' tweets after the change (concept drift) happens.

6. Conclusion

While there are many studies in literature that focused on sentiment analysis, social media mining, concept drift in machine learning, and text mining but however, previous research has not adequately focused on understanding the innovational change caused by concept drift while performing sentiment analysis. A similar issue was raised by (Ozgun and Broekel 2021) indicating that, so far, previous research has not adequately addressed the potential variations in news media's content and sentiment with respect to technologies.

Concept Drift is a phenomenon associated with classification models, especially in dynamic environments, which makes the model's prediction not sufficient anymore. Many studies in literature have focused on drift detection methods, systems, and corrective approaches, but limited studies focused on analyzing and/or understanding the change caused by concept drift over time, especially in the field of new inventions or modern technologies. Most of the studies refers to concept drift as negative effect, since the evaluation is from model accuracy perspective, but the change resulted by this phenomenon in technology field is not necessary to be negative, since it may lead to a new products or new inventions, and this was the main argument of this study.

The tweets data was preprocessed, and sentiment analysis was done on the processed tweets, showing that the large segment of technology companies' tweets is positive. Then the data was used to make predictions using the two machine learning algorithms which are:

- Naive Bayes Classifier
- XGBoost Classifier

The accuracies attained by both the Naive Bayes Classifier and the XGBoost Classifier are ~73% and ~82% respectively, but these accuracies fall to the ~30-40% mark when shown new data, which indicates that the label is drifting rapidly with respect to the new data. To visualize this, the line plots were observed which show a sudden drop in model performance and therefore lead to the conclusion that sudden concept drift is present in this use case.

The findings of this study show that machine learning models can be used in many fields to analyze and understand the public opinion about specific service or products, and the effects of sudden or gradual changes on that opinion, and this is particularly important for decision support systems in marketing, market analysis, politics, and social studies. Furthermore, these machine learning tools can cooperate with artificial intelligence systems to form most effective decision support systems, especially since the incorporation of AI and ML into products and services involve

substantial innovation by firms. Yet, we still lack insight into how these innovations are communicated (Fredstrom, et al. 2022).

This study tried to spot the light on understanding the changes that happens in technology through machine learning tools and introduce new reading to these changes as innovative change, which may lead to new products or technologies. Such an approach can provide important tools for technology companies, market analysts and researchers in this field to get an insightful forecast of the trends of future technologies and products.

7. Future Work:

This study presented an approach to understand the change in technology (innovation) based on tweets. The study focused on 270K tweets of ten technology companies, the focus was on analyzing the sentiment of these tweets and detecting/identifying the concept drift that occurred and studying the results of this drift / change to predict the pattern of results that could cause the deviation of the concept, which can in turn give some insights about future technological change or innovation.

The future work may include one or more of the following:

1. To use wider data set by increasing the number of tweets and the number of companies to expand the data set and include some accounts that usually focus on technology news and latest updates and inventions.
2. In our study, we depended on one commonly used algorithm for text classification (Naïve Bais) and another state of art algorithm in the field, which is XGBoost classifier, it worth to include other algorithms such as artificial neural network (ANN), Random Forests and Support Vector Machine (SVM), to compare the accuracy of the results and choose the best performance algorithm(s).
3. In our study, we chose the Error rate drift detection method, and there are many other similar methods that can be used for drift detection. Future work may also consider using different methods and comparing the results.
4. Lastly, Re-perform the study depending on the Patent Records/Documents as new data set for the research and compare the results with tweets to better understand the technology changes or innovation.

References

Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums." *ACM Transactions on Information Systems* 1-34. 2008.

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. "Sentiment analysis of twitter data." *Proceedings of the workshop on languages in social media.* Portland, Oregon: Association for Computational Linguistics. 30-38. 2011.

Alencia, Calandra , Achmad Nizar, Nur Fitriah Ayuning, and Herkules. "Sentiment Analysis of Online Auction Service Quality on Twitter Data: A case of E-Bay." *The 6th International Conference on Cyber and IT Service Management (CITSM 2018).* Medan, Indonesia. 2018.

Baena-Garcıa, M., J. del Campo-Avila, A. Bifet R. Fidalgo, R. Gavalda, and R. Morales-Bueno. "Early drift detection method." *4th Int. Workshop Knowledge Discovery from Data Streams.* 2006.

Barzenji, Hawar Sameen Ali. "Sentiment Analysis of Twitter Texts Using Machine Learning Algorithms." *Academic Platform Journal of Engineering and Science* 9 (3): 461-471. 2021.

Bayram, Firas, Bestoun S. Ahmed, and Andreas Kassler. "From concept drift to model degradation: An overview on performance-aware drift detectors." *Knowledge-Based Systems .* 2022.

Bhuvaneswari , Muthukumar , and Vasudevan Srividhya. "Enhancing the Sentiment Classification Accuracy of Twitter Data using Machine Learning Algorithms." *CORE UK* Volume I (ISSN: 2349 – 4891): 1. 2017.

Bifet, Albert, and Ricard Gavaldà. "Learning from Time-Changing Data with Adaptive Windowing." *The 2007 SIAM International Conference on Data Mining (SDM).* 443 - 448. 2007.

Dave, Kushal , Steve Lawrence, and David M. Pennock. "Mining the Peanut Gallery- Opinion Extraction and Semantic Classification of Product Reviews." *Proceedings of the 12th International Conference on World Wide Web,* 2003. *WWW 2003.* Budapest, Hungary.

Devi, B. Lakshmi, V. Varaswathi Bai, and Somula Ramasub. "Sentiment Analysis on Movie Reviews." In *Emerging Research in Data Engineering Systems and Computer Communications*, 321-328. 2020. India: Springer Nature Singapore Pte Ltd.

Elwel, Ryan, and Robi Polikar. "Incremental Learning of Concept Drift in Nonstationary Environments." *IEEE Transactions on Neural Networks* 22 (10): 1517-1531. 2011.

Fredstrom, Ashkan, Vinit Parida, Joakim Wincent, David Sjodin, and Pejvak Oghazi. "What is the Market Value of Artificial Intelligence and Machine Learning? The Role of Innovativeness and Collaboration for Performance." *Technological Forecasting & Social Change* 180 (121716). 2022.

GAMA, JOAO , INDRE ZLIOBAITE, ALBERT BIFET, MYKOLA PECHENIZKIY, and ABDELHAMID BOUCHACHIA. "A Survey on Concept Drift Adaptation." *ACM Computing Surveys* 1 (1): 35. 2014.

Gama, João, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. "Learning with Drift Detection." In *Advances in Artificial Intelligence – SBIA 2004*, 286–295. 2004. Berlin, Heidelberg: Springer.

Gangawane, A.A, and H. B. Torvi. "Opinion Mining and Sentiment Analysis on Twitter." (International Journal of Innovative Research in Science, Engineering and Technology) 6 (7). 2017.

Gemaq, Rosana, Albert F J Costa, Rafael Giusti, and Eulanda Santos. 2020. "An overview of unsupervised drift detection methods." *WIREs Data Mining And Knowledge Discovery by Wiley Periodicals LLC* 10 (6).

Gözüaçık, Ömer, and Fazli Can. "Concept learning using one-class classifiers for implicit drift detection in evolving data streams." *Artificial Intelligence Review* 3725–3747. 2021.

Gu, Feng, Guangquan Zhang, Jie Lu, and Chin-Teng Lin. "Concept drift detection based on equal density estimation." *International Joint Conference on Neural Networks (IJCNN).* Vancouver, BC. 2019.

Gupta, Ankita, Jyotika Pruthi, and Neha Sahu. "Sentiment Analysis of Tweets using Machine Learning Approach." *International Journal of Computer Science and Mobile Computing* 6 (4): 444 – 458. 2017.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining- Concepts and Techniques.* Elsevier Inc. 2012.

Hu, Minging, and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* 168. 2004.

Janardan, Shikha Mehta. "Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues." *5th International Conference on Information Technology and Quantitative Management (ITQM2017).* India: Faculty of Mathematics and Informatics, Vilnius University,. 2017.

Kantardzic, Mehmed. *DATA MINING Concepts, Mode, Methods, and Algorithms.* Vol. THIRD EDITION. New Jersey: John Wiley & Sons, Inc., Hoboken. 2020.

Li, Ping. "Robust LogitBoost and Adaptive Base Class (ABC) LogitBoost." *Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010).* 2010.

Lu, Jie, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. "Learning under Concept Drift: A Review." *IEEE Transactions on Knowledge and Data Engineering* 31 (12): 2346 - 2363. 2019.

Lu, Ning, Guangquan Zhang, and Jie Lu. "Concep tdrift detection via competence models." *Artificial Intelligence* 11 (24): 2007..

M, Vijayalakshmi. 2015. "Identifying Concept-Drift in Twitter Streams." *International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).* Mumbai, India.

Maciel, Bruno Iran Ferreira, Silas Garrido Teixeira Carvalho Santos, and Roberto Souto Maior Barros. 2015. "A Lightweight Concept Drift Detection Ensemble." *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI).* Vietri sul Mare, Italy.

Ogul, Hasan, and Emrah Ekmekciler. "Two-way collaborative filtering on semantically enhanced movie ratings." *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces.* Cavtat, Croatia. 2012.

Ozgun, Burcu, and Tom Broekel. "The geography of innovation and technology news - An empirical study of the German news media." *Technological Forecasting & Social Change* 167 (120692). 2021.

Qahtan, Abdulhakim A., Basma Alharbi, Suojin Wang, and Xiangliang Zhang. "A PCA-Based Change Detection Framework for Multidimensional Data Streams: Change Detection in Multidimensional Data Streams." *The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 935–944. 2015.

Rianto, Achmad Benny Mutiara, Eri Prasetyo Wibowo, and Paulus Insap Santosa. "Improving the accuracy of text classifcation using stemming method, a case of non-formal Indonesian conversation." *Journal of Big Data* 8 (26). 2021.

Saberi, Bilal , and Saidah Saad. "Sentiment Analysis or Opinion Mining: A Review." *International Journal on Advanced Science, Engineering and Information Technology* Vol.7 (2017) No. 5 (2088-5334): 1. 2017.

Saberi, Bilal, and Saidah Saad. "Sentiment Analysis or Opinion Mining: A Review." *International Journal on Advanced Science, Engineering and Information Technology* Vol.7 No. 5. 2017.

Schuh, Günther, Gunther Reinhart, Jan-Philip Prote, Frederick Sauermann, Julia Horsthofer , Florian Oppolzer, and Dino Knoll. "Data Mining Definitions and Applications for the Management of Production Complexity." *52nd CIRP Conference on Manufacturing Systems* 1. 2019.

Tianqi Chen, Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016.

Vasudevan, Srividhya. "Enhancing the Sentiment Classification Accuracy of Twitter Data using Machine Learning Algorithms." In *Statistical Approaches in Multidisciplinary Research*, Chapter 21. India: SHANLAX PUBLISHER. 2017.

Wang, Heng, and Zubin Abraham. "Concept Drift Detection for Streaming Data." *International Joint Conference of Neural Networks 2015.*

Wang, Jenq-Haur, and Ting-Wei Liu. "Improving sentiment rating of movie review comments for recommendation." *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW).* Taiwan .

Widmer, Gerhard, and Miroslav Kubat. "Learning in the presence of concept drift and hidden contexts." *Machine Learning* 23: 69-101. 1996.

Yang, Lingkai, Sally McClean, Mark Donnelly, Kevin Burke, and Kashaf Khan. "Detecting and Responding to Concept Drift in Business Processes." *Algorithms* 15 (174): 1. 2022.

Zhou, Xujuan , Xiaohui Tao, Jianming Yong, and Zhenyu Yang.. "Sentiment Analysis on Tweets for Social Events." *IEEE 17th International Conference on Computer Supported Cooperative Work in Design.* Whistler, BC, Canada. 2013

Biographies

Alptekin DURMUSOGLU is an Associate Professor at Industrial Engineering Department at University of Gaziantep, Turkey. He holds a MSc and PhD degree in Industrial Engineering from the University of Gaziantep in Turkey. Dr. DURMUSOGLU's research examines the data mining and knowledge discovery. In addition, he has a long-standing interest in technology management with an emphasis on early warning for technological changes. Dr. Alptekin has extensive tenure in academia. He was an assistant professor at Gaziantep University, Turkey from 2012 to 2018 and he was teaching in Opole university of Technology, Poland within ERASMUS+ program in 2016 and in Lucian Blaga University of Sibiu, Romania, 2019. In addition to his teaching experience, Dr. Alptekin has held several management positions: vice dean, Gaziantep University, Faculty of Engineering, (2016-2021), member of intellectual property committee, Gaziantep University, (2017-2020), member of guidance committee, Turkish patent and trademark office, Gaziantep university, (2016-2020), member of board of directors, target technology transfer office, (2013-2020), head of industrial engineering division (ABD) of Ind. Eng. Dept, Gaziantep University, (2016-2019), and Member of Board of Directors, Gaziantep University, Graduate School of Natural and Applied Science, (2013-2019). Durmusoglu is a board member of Journal of IEEE Transactions on Engineering Management (2018-) and a Board Member of International Journal of Innovation and Technology Management (2017-)

Mohamad NACI, M.Sc. student at Industrial Engineering Department at University of Gaziantep. He held a bachelor's degree in control and automation engineering from the faculty of Electrical engineering at Aleppo University, Syria. Mohamad is an experienced engineer in electrical and electronic systems design and maintenance, and he has worked for many international companies such as Lafarge Cement in the period between 2010 – 2015. Mohamad has finished the first year of the master's degree in Control Engineering at Aleppo University in 2011, but due to the war in Syria he moved to Turkey and re-joined the department of Industrial engineering at Gaziantep university in 2019. During the period of the ongoing war in Syria, Mohamad volunteered to work in many humanitarian organizations to provide aid and relief to the Syrian people, and he was honoured at the International Conference on Syrian Education, which was held in Istanbul, Turkey in 2017.