EVALUATING DANCE SPORT FOR EXPERTISE: A CLASSIFICATION APPROACH BASED ON VERBAL DESCRIPTIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ALİ CAN SERHAN YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

NOVEMBER 2021

A BINARY CLASSIFICATION MODEL USE ON DANCE SPORT EVALUATION TO DETECT EXPERT & NOVICE

Submitted by ALİ CAN SERHAN YILMAZ in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Director, **Graduate School of Informatics** _____

Dr. Ceyhan Temürcü
Head of Department, **Cognitive Science** _____

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, **Cognitive Science Dept., METU** _____

Examining Committee Members:

Asst. Prof. Dr. Murat Perit Çakır
Cognitive Science Dept., METU _____

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science Dept., METU _____

Asst. Prof. Dr. Erol Özçelik
Psychology, Çankaya University _____

Date: *25.11.2021*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : ALİ CAN SERHAN YILMAZ

Signature : _____

ABSTRACT

EVALUATING DANCE SPORT FOR EXPERTISE: A CLASSIFICATION APPROACH BASED ON VERBAL DESCRIPTIONS

Yılmaz, Ali Can Serhan

M.Sc., Department of Cognitive Sciences

Supervisor: Assoc. Prof. Dr. Cengiz Acartürk

November 2021, 74 pages

Dance sport is a performance in which a male and a female aim at exhibiting certain body techniques in a limited time period. The judges, who report the performance evaluation, focus on the body lines, the bodily communication between the partners, and the sense of rhythm. The evaluation process of dance sport may vary across judges' level of experience. The goal of the thesis is to investigate the modeling capabilities of dance performance evaluation, which eventually may lead to a binary classification of expert and novice evaluators through their verbal descriptions of the dance activity. In particular, the models presented in this thesis aim to classify the evaluator as an expert or a novice through the analysis of speech data. For this, we trained two binary classification models, namely Multinomial Naïve Bayes and DistilBERT. The findings reveal that both models may return acceptable results for English and Turkish, though their different performance in accuracy.

Keywords: Binary Classification, Natural Language Processing, Dance Performance , Multinomial Naïve Bayes, DistilBERT.

ÖZ

DANS SPORU DEĞERLENDİRMESİNDE UZMANLIK: SÖZEL BETİMLEMELERE DAYALI BİR SINIFLANDIRMA YAKLAŞIMI

Yılmaz, Ali Can Serhan

M.Sc., Bilişsel Bilimler

Danışman: Doç. Dr. Cengiz Acartürk

Kasım 2021, 74 sayfa

Dans sporu, erkek ve kadının birlikte belirli vücut tekniklerini kısıtlı bir zamanda müzik ile beraber uyguladığı bir performans faaliyetidir. Hakemler çiftleri değerlendirirken partnerler arası iletişime, vücut çizgilerine, çiftlerin müzik ve ritim ile uyumlarına odaklanmaktadır. Bu performansın değerlendirilme şekli değerlendirecek kişinin tecrübe seviyesine göre değişiklik göstermektedir. Bu tez çalışmasının amacı dans sporu değerlendirilmesinin, uzman ve tecrübesiz değerlendiricilerin sözel verilerini kullanarak ikili sınıflandırma ile modellenebilirlik kapasitesini araştırmaktır. Bu tez çalışmasında sunulan modellerle değerlendiricilerin sözel verilerini analiz ederek onları uzman veya tecrübesiz olarak sınıflandırmak amaçlanmaktadır. Bunu yapabilmek için Multinomial ve DistilBERT olmak üzere iki ayrı ikili sınıflandırma modeli kullanılmıştır. Sonuçlar gösteriyor ki, kullandığımız iki model de hem İngilizce hem de Türkçe veri setleri için kabul edilebilir doğruluk oranları vermektedir

Anahtar Sözcükler:Doğal Dil İşleme, İkili Sınıflandırma, Dans Performansı, Multinomial Naïve Bayes, DistilBERT.

To My Mom and Grandma

ACKNOWLEDGMENTS

First and foremost, I want to convey my thanks and respect to Assoc. Prof. Dr. Cengiz Acartürk, my thesis advisor, for his invaluable assistance, encouragement, and continuous support throughout this research.

I like to thank the members of my thesis jury, Asst. Prof. Dr. Murat Perit Çakır and Asst. Prof. Dr. Erol Özçelik for their feedback and ideas.

I also like to thank Fulya Gökalp Yavuz and Deniz Demirci, who supported me with their suggestions and ideas along my new machine learning journey.

Finally, I would like to express my deepest gratitude to my sister Ceylan, my parents Neslihan & Bektaş, my grandma Yıldız and my partner Elgin for their love, encouragement, and support. This thesis would not have been accomplished without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BPM	Beat Per Minute
CP	Choreography & Presentation
DistilBERT	Distilled version of Bidirectional Encoder Representations from Transformers
JS	Judging System
MM	Movement to Music
MNB	Multinomial Naïve Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PB	Posture Balance Configuration
PS	Partnering Skill
QM	Quality of Movement
SA	Situation Awareness
SP	SelectPercentile
Tf-Idf	Term Frequency / Inverse Document Frequency
TQ	Technical Quality
WDC	World Dance Council
WDSF	World Dance Sport Federation

CHAPTER 1

INTRODUCTION

1.1. Motivation

Numerous research methodologies have been employed to perform a comparative analysis of expert and novice performance in various domains. The analyses relevant to Cognitive Science include teaching (Hogan et al., 2003), reading (Peskin, 1998), understanding a complex system (Hmelo-Silver & Pfeffer, 2004), decision making (Phillips et al., 2004), game playing (Engle & Bukstel, 1978), among many others. In the present study, we focus on a comparative analysis of expert and novice evaluators for dance sports by paying special attention to the verbal utterances of the judges who evaluate dance performance.

Dance sport can be analyzed from various perspectives, and based on the analysis assumptions, various definitions may be made to identify their salient aspects. For instance, Dan Năstase (2012) reviews the possible definitions of dance sport considering approaches from spatial, pedagogical, social, and structural aspects. In summary, they define dance sport as "… an art-sport that originates in the social couple's dances based on a time-limited complex motion activity and as execution rhythm, by a melody, and spatially by a dance floor [sic.]." (p. 3). Since dance sport is a very complex activity, it demands so much about cognition. Thus, the analysis of its evaluation process can open new doors to the way experts use their abilities related to cognition while judging dance couples. This is the main motivation of the present study.

Traditionally, the evaluation of dance sport performance is conducted by former experienced dance sport competitors. Starting from the first round to the semi-final round of a competition, the adjudicators pick the couples who they think as qualified enough for the next round. At the final round, depending on the type of the competition, adjudicators either rank the six couples from 1 to 6 or rate the couples by assigning scores from 1 to 10 in four different criteria, namely *technical qualities* (TQ), *movement to music* (MM), *partnering skill* (PS), and *choreography & presentation* (CP).

Regardless of the level of experience the judges have, their opinions about couples are represented by integers in skating reports[1]. The specific motivation of the present study is to investigate the underpinnings of experts' and novices' decision-making processes and to study the capability of computational models that are able to detect regularities behind the judging processes in terms of NLP (Natural Language Processing) analyses.

[1] The report that presents all markings of the adjudicators in a dance sport competition is called *skating report*.

1.2. Research Question

In this study, we hypothesize that it is possible to develop a computational modeling approach by analyzing linguistic utterances from expert and novice evaluators such that the model achieves higher than the chance factor. To test this hypothesis, we have used Multinomial Naïve Bayes (MNB) Classifier (Tf-Idf) and DistilBERT classifier. We collected speech data in Turkish from 22 participants and collected text data in English from online resources, such as Reddit and Youtube. We shaped our datasets in two different ways: Balanced (equal number of words in every row) and Sentence-fragmented (one sentence per row), and we compared the outputs of DistilBERT and MNB. We assumed that the novice participants did not have any experience in dance performance evaluation. Therefore, this study conceived as the analysis of expertise in evaluation, too.

Although our main focus was on developing a model that can offer an insight into experts' way of combining the knowledge to evaluate couples on the dance floor, we also wanted to see how novices yield their attention while trying to make an evaluation on a task that they have not enough knowledge. We took notes of the noticeable differences between experts and novices for the Turkish dataset. Also, we benefited from feature extraction tools to see what our model could capture as well.

1.3. Scope of the Thesis

We used unsupervised machine learning to categorize the Turkish dataset and supervised machine learning for both datasets for binary classification. MNB and DistilBERT models were employed, and the results were compared for both datasets with all the model versions. `Max-features`, `max_df`, and `SelectPercentile` parameters were manipulated in each run on the MNB classifier, while the `learning rate` was the parameter that was changed in the runs of the DistilBERT classifier.

The Turkish dataset was established by converting the speech data into text. Ten experienced dance sport Latin branch performers and 12 novice participants, who have no experience in dance sports or minimal experience in other dance-related sports, were shown a video of the final round of the WDSF Junior 2 Latin World Championship[2]. The Turkish dataset was categorized as expert and novice based on clustering answers of participants to 10 Likert questions. The English dataset was collected from the written comments made to competition videos of a dance sport athlete on the DanceSport channel of reddit.com[3]. The comments were classified as either the "intermediate level of dancing experience" or the "highest level of dancing experience" in case the transcript belongs to a former world champion, present adjudicator, or a dance teacher. This categorization was performed manually, as an operational assumption at the level of the dataset, at the present study.

[2] This is a yearly organized competition by World Dance Sport Federation (WDSF) for the dancers aged between 14-15. For more detailed information, https://www.worlddancesport.org/Event/Competition/World_Championship-Moscow-19362/Junior_II-Latin-45081 may be visited.

[3] https://www.reddit.com/r/DanceSport/comments/8j1qj0/critique_for_gold_amam_latin/. Retrieved on November 2, 2021.

1.4. Thesis Organization

There are five chapters in this study. Chapter 2 presents a background and literature survey for relevant expert-novice studies in Cognitive Science as well as background information about natural language processing. We presented our methodology in Chapter 3. The results of the study are reported, and the findings are discussed in Chapter 4. In Chapter 5, we conclude our study by proposing limitations and future work.

CHAPTER 2

2. BACKGROUND & LITERATURE SURVEY

Our study takes place where dance, evaluation of sports, expert-novice paradigm, and text classification models emerge. Hence, previous research on these concepts was analyzed while running the literature survey. However, since we could not find any other research employing these four concepts at the same time, we gathered information from the studies that have either one of these concepts or a binary combination of these four. Although the tasks in which expert-novice differentiation is analyzed were the focus of the literature review, some of the studies that run text classification models to distinguish the relevant data were also mentioned.

2.1. Research on Sports in Cognitive Science

Perceptual and cognitive tasks like attention, visual discrimination, problem-solving, prediction, and judgment are the foundation for the capabilities required in many sports activities(Mann et al., 2007). For instance, athletes should be able to shift their focus to the most beneficial locations, a specific portion of a body, for instance, in order to select and derive necessary information from the surroundings (Russo & Ottoboni, 2019).

As sports is a field in which cognitive capabilities are highly demanded, there has been increased research interest in perceptual and cognitive aspects of athletes in the past few decades. In this chapter, we review the studies on the role of expertise in sports.

The ability to continuously display better athletic performance has been termed as sport expertise. Despite the fact that higher performance can be observed at first glance, the perceptual-cognitive factors that lead to the expert advantage are less visible (Mann et al., 2007). Mann and his colleagues suggest that it is very important to know where and when to look to achieve a good sports performance. However, sportsmen are usually synchronously overloaded by irrelevant visual information, as well as useful information during a performance. Mann et al. questioned whether there was a statistically significant difference between experts and non-experts, both in performance and gaze behavior during task performance. Sport types were taken into hand in 3 categories, namely interceptive sports (sports that are done by a held implement and require an object control in environment, such as squash, badminton, tennis), strategic sports (sports that involve multiple teammates with a diverse placement going back and forth between defense and attack, such as field hockey, soccer), and other sports (self-paced and aim-oriented sports, such as billiards, golf, target shooting). The results showed that experts were more accurate in their decision-making than their less experienced colleagues in terms of the examination of performance measures. Also, experts predicted their competitors' intents significantly faster than less proficient participants. Even though experienced athletes were observed to be more precise than the novice athletes across the three sport types included in the study, the degree of the difference was largely stable, implying that response precision was not affected by the type of sport.

To provide evidence of how expertise influences brain functionality, researchers have spent recent years exploring whether sports practice may increase abilities in both sportive and general domains (Debarnot et al., 2014). Cognitive-perceptual skills such as anticipation, judgment, and context awareness are requirements for outstanding performance in athletics. Experienced athletes build complex task-specific knowledge structures as a result of extensive practice, allowing them to cope with problems more quickly and successfully than others (Pio Di Tore, Alfredo(University of Salerno & Raiola, Gaetano(MIUR Campania, 2018). Situation awareness (SA) is one of these cognitive-perceptual skills that are worth to be analyzed in sports. According to Endsley, SA is "the perception of environmental factors within a volume of time and space, the understanding of their significance, and the projection of their state in the near future" (Endsley, 1988). Raiola suggests that experienced sportsmen's SA rely on simultaneous processing, which picks meaningful data from a plethora of available data instead of using linear pre-processing-action systems (Raiola, 2017). This supports the idea that experience enables elite athletes to cope with problems more quickly than less experienced athletes.

2.1.1. The Role of Expertise in Dance & Sports

Matthew Kenney Henley investigates the differences between expert and novice dancers by testing them with a questionnaire to understand if participants recall Shape, Space, Time, and Effort elements after showing them a short contemporary dance video (Henley, 2014).

While collecting the data, we asked Likert questions to the participants based on their experience level on dancing. We noted the answers and tried to cluster the data into two parts. As we found out there is a significant difference between the two groups of data, we could categorize the participants as novices and experts.

Henley's study does not really show a statistically significant difference between expert and novice dancers on Shape, Time, and Effort elements, yet the results indicate that experts are better at recalling the phrases overall. Statistically significant difference occurs at the Space element. Expert dancers can connect spatial elements of a dance phrase to other elements to achieve a better recall performance. This can be considered as chunking of spatial elements, such as the direction the dancer was looking or the angle the dancer moved.

A dance choreography can be considered as a set of movements that match with certain music on a certain path on the dance floor (Ofli et al., 2012). While dancing, the music provides constant sensory inputs that are synchronized with the activity and allows for a lot of body movement (Merom et al., 2013). Learning a choreography requires visual attention, muscle control, and music & rhythm awareness. As dancing demands so much about cognition, it is no surprise that learning or performing dance has been a hot topic of research done by the Cognitive Science Faculty in recent years. (Carey et al., 2019) collected data to determine the impact of dancers' experience on motor imagery (MI; proceeding actions only mentally without actually moving) and attentional effort (measured by tracking the movements of the pupil) while learning, performing, and envisioning a dance routine.

In this study, 18 female ballet and dancers who have three different experience levels (novice, intermediate, expert) were selected to be shown a piece of choreography that lasts 15 seconds. Data were collected while these dancers learned, performed, and imagined the choreography. In order to examine the imagery abilities of the participants, two MI questionnaires were given to participants. Tobii eye-tracking glasses were employed to track participants' eye movements. The time spent for performing and imagining the choreography was measured via a stopwatch so that relevant analyses could be done.

Results indicated that experts and novices do not differ significantly, which means dance expertise does not affect MI significantly. However, since novice dancers have the highest and intermediate dancers

have the lowest pupil dilation at the beginning of the choreography, it can be claimed that these dancers differ on attentional effort.

In our study, we used a similar categorization method in two different settings. The first, data collection from experts and novices via a data collection, and second, data collecting from intermediate level experienced vs. much higher level experienced via a web search.

2.1.2. Expert-Novice Differentiation in Judgement Process

In some sports, such as football, basketball, handball, and many others, judges' mission is making sure that players perform within the rules defined for that particular sport and involve in the competition when rules are broken. However, there are some other sports in which the winner is selected directly by judges, such as figure skating, gymnastics, dancesport, etc.

The evaluation of Gymnastics is a complex process since it requires action in a limited time when there are limited resources (Mercier & Heiniger, 2018). In gymnastics, the judge's job is to watch a very quick demonstration and make a judgment depending on how that presentation is perceived (Ste-Marie & Lee, 1991).

A single earlier exposure to a term has consequences on memory-influenced decisions, both conscious and unconscious (Ste-Marie & Lee, 1991). Memory' for previous events has also been demonstrated to impact subjective assessments (Jacoby et al., 1988). As mentioned in (Ste-Marie & Lee, 1991), an athlete is given a quick warm-up before a qualifying performance in a standard gymnastics' tournament. The athlete will usually practice warm-up the same routine that will be judged in the competition. It should be noted that judges watch both the qualifying performance and warm-up. Ste-Marie and Timothy D. Lee (1991) come up with an example of an athlete who makes a mistake while warming up. Judges may be biased and shift the grounds of their perception even if the athlete's competition performance is much better. It is stated that by being more conscious of how a previous event influences our current decisions, we can shift the grounds of our judgments and avoid the consequences of that event (Jacoby & Kelley, 1987). Even we mention that experts might be biased, there are some studies that show experts are less biased in the topics that they have expertise. Accountants were less prejudiced on accounting judgments than non-accounting judgments, according to Smith and Kida (1991). Expertise, without a doubt, provides significantly more value to decision-makers than bias elimination(K. Phillips et al., 2004).

Melanie Mack (2020) run a study to investigate cognitive and perceptual evaluation practices in Gymnastics by using critical kinematics information. In this study, eye tracking and performance evaluation via six-point Likert questions were implemented on participants from different levels of expertise in Gymnastics (Mack, 2020). The number of fixations, average fixation duration, summarized fixation duration, judgment score, and judgment accuracy of the participants were analyzed. Participants were categorized as visual-motor experts and novices. The results do not show a significant difference in gaze behavior. Nevertheless, there is a significant difference between the groups in judging accuracy. Surprisingly, the group of novice people performs better than motor experts on judging accuracy. This shows that the relationships between level of expertise & judging and level of expertise & gaze behavior are not going linear.

In another study run on expert-novice differences in gymnastics judging, twelve novice and expert judges were assessed on different domains (Ste-Marie, 1999). Judges that have ten years or more experience and have higher than Level V certificate from the Ontario gymnastic judging system are considered as experts, whereas judges that have three years or less experience and certified at Level 1 or 2 are taken into account as novices.

3-5 gymnastic figures are combined and performed on bars, beam, and floor. Twenty-four gymnastic sequences were selected from these figures and formatted onto a videotape. These gymnastic sequences are divided into four blocks, including equally distributed six sequences that have the same number of different events. The sequences were shown in a fashion such that approximately three seconds of footage of gymnastic routines were shown part by part to participants, and after every part, participants were asked questions that measure if they could predict the next figure. The study also seeks the answers to the questions of whether a better prediction of the next figure also results in a more accurate evaluation.

Results indicated that expert judges are better at predicting the next gymnastic element when compared to novice judges. Also, when the depth of knowledge was observed for the gymnastic elements as identification, symbol code, and level of difficulty, it was seen that expert judges were significantly more accurate than novice judges in giving information about the symbol code and the level of difficulty. This shows that better prediction skills result in better accuracy at evaluation. (Poulton, 1957) mentions that to achieve a precise acquisition of a moving target, one must know ahead of time where the target will be after his reaction movement is completed and must have information about the environment where movement occurs. Ste-Marie (1999) suggests the reason better prediction skills result in better accuracy at evaluation might be that successful prediction of the next gymnastic element may enable the judge to retrieve the information needed for that particular gymnastic element right before it occurs. To be able to access that information beforehand may provide a guideline to evaluate while watching the performance.

In this study, we used Dance Sport Evaluation as a task to observe differences in speech data between experts and novices. Subjective evaluation is still a problem for the competitions of both WDSF and WDC in the national and international settings (Gürbüz, 2018). Thus, there is a need for a judging system that enables judges to evaluate couples objectively. There is a study that investigates possible ways to eliminate subjectivity in the evaluation process in dance competitions (Silvia & Alexandru Adrian, 2016).

In their study, Silvia Teoderescu and Adrian Nicoara (2016) try to find out the criteria that can be implemented to increase objectivity by collecting data from a questionnaire including 26 items which draws the ideas of Romanian judges recognized nationally and internationally.

Results indicated that 60% of judges think that the current system is needed to be changed. Also, 60% of judges pointed out that every other criterion should be evaluated by a judge so that judges can use the time efficiently by focusing on only one aspect on the dance floor. Even though judges agreed on the time given for each dance, which is about 1.5-2 minutes, is enough for the couples to show their level of technique, musicality, and artistic capabilities, the number of couples in each heat (A heat indicates the couples on the floor at the same time. At first rounds, a heat consists of averagely 8-9 couples. Starting from the quarter-finals a heat includes six couples) should be decreased to increase time spent per couple to be evaluated in more detail. Another idea agreed by the majority is that the artistic part and technical part should be evaluated separately while considering extra points for more difficult figures. 90% of the participants agreed that dance sport should be included in Olympic Games. However, they believe that it is possible if only subjectivity can be removed.

The presented thesis study may contribute to finding out the aspects that indicate subjectivity in dance sport evaluation by demonstrating the reasoning of adjudicators underneath their ratings. Information retrieval methods can be employed to reveal exactly where adjudicators focus on while evaluating and what is needed to be improved.

2.2. Dance Sport as a Sport Branch

Dance sport is an activity in which male and female dancers try to combine sport and dance by demonstrating certain techniques while staying with the rhythm (that matches with the required BPM for the relevant dance) within a limited amount of time. In general, dance sport is classified into two groups (Latin and Standard). Each group consists of five dances: Samba, Cha Cha Cha, Rumba, Paso Doble, and Jive in Latin, and Foxtrot, Slow Waltz, Viennese Waltz, Tango, Quickstep (Standard). This classification has been proposed by two recognized federations, namely World Dance Sport Federation (WDSF, granted for full recognition by International Olympic Committee) and World Dance Council (WDC). Dancesport figures are listed in the syllabus published by the WDSF. There are 233 figures listed under the Latin section of WDSF, and all the figures are explained in terms of a set of features, such as the start and finish positions, timing, couple positions, and notes with use alternatives. Figure 1 shows an example specification of two dances. These figures are also known as basic figures.

Table 2. 1 Example of figures listed in WDSF Syllabus

OPEN HIP TWIST	CLOSE HIP TWIST
Start: LF fwd, T turned out (Open Opp. LH to RH)	**Start**: LF fwd T turned out (Close Opp.; Normal Hold)
Finish: RF to side (Fan L Angle; LH to RH)	**Finish**: RF to side (Fan L Angle; LH to RH)
Timing: 2 3 4&1 2 3 4&1	**Timing**: 2 3 4&1 2 3 4&1
NOTE - General: Steps 1-5 or 6 -10 only may be used.	**NOTE** - General: Steps 1-5 or 6-10 only may be used.
NOTE -Couple Position: May be danced in Open Opposing position	**NOTE** - General Action: Steps 8-10 may be replaced by a Cha Cha Lock fwd or three Cha Cha Locks fwd (Man) and a Cha Cha Lock bwd or three Cha Cha Locks bwd (Lady). When steps 1-5 only are used, Lady may dance a Cha Cha Chasse on steps 3-5 turning L to end in Close Opp. Pos.
NOTE - Timing: Guapacha Timing may be used.	**NOTE** - Lead/ Hold/ Shaping: It may be danced with RH-to-RH Hold, changing to LH to RH hold on step 7.
NOTE - General Action/ Couple Position: The figure may be used as a Foot Change-in this case Lady will dance a Spiral Cross on step 7 and follow with Cha Cha Chasse to side to end in L Side Same Position. Man will replace steps 8-10 with a side Rock (RF, LF) timed 4 1 and release hold at the end.	**NOTE** - General Action/ Couple Position: The figure may be used as a Foot Change- in this case Lady will dance a Spiral Cross on step 7 and follow with Cha Cha Chasse to side to end in L Side Same Position. Man will replace steps 8-10 with Rock to side (RF, LF) timed 4 1 and release hold at the end.

In dance sport, the quality of the basic figures determines the quality of a couple. As couples gain experience, their choreographies get more and more complex with trending moves.

However, no matter how much they enrich their dancing with free style figures, artistic moves and different styles, the evaluation process is mostly done by analyzing the quality of the basic figures of couples as listed in WDSF syllabus. Experienced dance sport athletes are the ones that master the basic figures of dance sport.

2.3. The History of Judging Systems Implemented by WDSF

WDSF has been using different judging systems for the competitions of different segments in recent years. There are six different types of competitions hosted by WDSF, namely World Championship, Continental Championships, Grand Slam, World Open, International Open, and Open competitions.

The couples registered to WDSF are listed based on their world ranking. The summation of the highest six competition points within a competition season determines the couple's ranking. The points gained from a competition depend on the couple's ranking in that competition and the competition coefficient. The competition coefficient is directly related to the segmentation of the competition. The highest coefficient belongs to Grand Slam competitions, and it is followed by World Open competitions. Continental Championships and World Championships do not bring any points to the couples.

The judging systems that have been used within the last 12 years are The Skating System, Judging System 1.0, Judging System 2.0, Judging System 2.1, and Judging System 3.0. As explained before, the correct technique of the figures in dance sport has been explained in WDSF sources. Couples are evaluated on how much they can get close to the correct technique, how they interpret music & rhythm with their dancing and the artistic capacity that they can show on the dance floor. These three main aspects, namely technique, music & rhythm, and artistic capacity, have been the main considerations of judges since the first-ever skating system. In the evaluation process, the aspects that are judged did not change; however, the way they are judged has been evolved throughout the dancesport history.

The skating system can be considered as a ranking system in which judges rank couples from 1 to 6 in the final round. In all WDSF competitions except Grand Slams, this system was used until 2017. In this system, judges mark the couples they want to see in the next round until the final round. Judges need to mark just as many as the Chairman requests. For instance, if there are 38 couples dancing, and Chairman asks for 24 couples in the quarter-final round, every judge needs to mark 24 couples during the election. The first 24 couples with the most marks along the five dances are selected for the quarter final. Table 2.2 represents an example of a ranking report. All of the judges want to see couple number 12 and 37 for the next round.

Table 2. 2 A Ranking Report Example (Judges: A, B, C, D, E, F, G | Couples: 9, 11, 12, 23, 25, 37, 43)

Couple No	A	B	C	D	E	F	G	Total
43	x	x		x	x		x	5
11		x	x	x	x	x	x	6
25	x			x				2
23		x						1

37	x	x	x	x	x	x	x	7
12	x	x	x	x	x	x	x	7
9			x	x	x			3

At the final round, judges rank the couples by assigning unique numbers for each couple between 1st and 6th. Judges cannot assign same ranking to more than one couple, and they cannot choose not to assign a ranking to a couple. Table 2.3 presents an example of final round ranking with five different judges and six couples. The rankings that have given by the judges are placed in 'Judges' column, and total rankings gained are presented in 'Places' column.

Table 2. 3 Skating Report of Final Round

	Judges					Places						
Couple No	A	B	C	D	E	1	1-2	1-3	1-4	1-5	1-6	Result
51	1	1	1	2	1	4						1
52	4	2	2	1	2	1	4					2
53	3	3	3	5	4			3				3
54	2	4	5	4	3		1	2	4			4
55	5	6	4	3	5			1	2	4		5
56	6	5	6	6	6					1	5	6

Natasha Ambroz, WDSF Education Chair, suggested that skating system has been used for very long time and although it is very practical and easy to use, it is needed to be improved (WDSM Jan. 2010). To take the judging system to a more objective and transparent setting, especially for the final round, New Judging System (JS 1.0) was employed in 2009 as mentioned in(*World DanceSport Federation, Judging System*, n.d.).

The concept and the rules of JS 1.0 was developed by Japan Dance Sport Federation (WDSM Jan. 2010). This was the first time that numerical evaluating system is employed instead of ranking the couples at the final round. Five criteria were selected to be evaluated numerically, starting from 1 (very poor) to 10 (outstanding). An example final round skating report of JS 1.0 is presented in Figure 2.1. In Figure 2.1 there are only two couples presented, couple number 1 and couple number 7. The judges are starting from A to K, 11 judges in total.

Solo Dance(Waltz)

No	Judge	PCS [PB] Posture, Balance, Coordination	PCS [QM] Quality of Movement	PCS [MM] Movement to music	PCS [PS] Partnering Skill	PCS [CP] Choreography and Presentation	SubTotal PCS	SubTotal PCS
1	A	8.50	8.50	9.00	8.50	8.50	43.00	43.89
	B	8.00	8.00	8.50	8.50	7.50	40.50	
	C	9.00	9.50	9.00	9.50	8.50	45.50	
	D	8.50	8.50	8.50	8.50	8.00	42.00	
	E	9.00	8.50	8.50	9.00	8.50	43.50	
	F	8.50	9.00	8.50	8.00	8.50	42.50	
	G	8.50	8.50	9.00	9.00	8.50	43.50	
	H	9.00	9.00	9.00	8.50	8.50	44.00	
	I	9.50	9.00	9.00	8.50	9.00	45.00	
	J	9.00	9.50	9.00	9.50	9.50	46.50	
	K	9.00	9.50	9.50	9.50	9.00	46.50	
	Decision	8.78	8.89	8.83	8.83	8.56		
7	A	9.00	9.50	9.50	9.50	9.50	47.00	46.17
	B	9.50	9.50	9.50	9.50	9.00	47.00	
	C	8.50	9.00	9.00	8.50	8.50	43.50	
	D	9.00	9.00	8.50	9.00	9.00	44.50	
	E	8.50	8.50	9.00	9.00	9.00	44.00	
	F	8.50	9.00	8.50	8.50	9.00	43.50	
	G	9.50	9.50	9.50	9.00	9.50	47.00	
	H	9.50	9.00	9.50	9.50	9.00	46.50	
	I	9.50	9.50	10.00	9.50	9.00	47.50	
	J	9.50	9.50	10.00	10.00	9.50	48.50	
	K	9.50	9.50	9.50	9.50	9.50	47.50	
	Decision	9.17	9.28	9.33	9.22	9.17		

Figure 2. 1 Final Round Report of Judging System 1.0 from International DanceSport Federation Grand Slam Competition held in Seoul 2011

The JS 1.0 has five criteria, namely posture balance coordination (PB), quality of movement (QM), movement to music (MM), partnering skill (PS), and choreography & presentation (CP). All judges evaluate couples in 5 criteria, and at the end, the couple who has the most points is granted for the 1st rank, and other couples are ranked in the same way respectively.

Odd numbers of judges were employed in competitions to prevent couples from having the same result. Even though JS 1.0 brought more sensitive and transparent measurement to the table, a more fair and objective system was yet to come. Since there are lots of aspects to be considered within the five criteria across five dances, finalists were needed to be observed one by one with solo performances. At the final round, there were six couples, and thus there were 30 solo performances to be evaluated separately. This implementation aimed to spend the effort and time evaluating all couples equally. Even though some of the adjudicators mentioned that JS 1.0 was successfully employed in the Grand Slam Final in Shanghai, some athletes pointed out that it takes too long to finish the competition, and couples should be dancing all together on the floor (WDSM Jan. 2010). The WDSF Education and Sports Departments worked from 2010 to 2012 to teach judges about the usage of the system and improve it further. Since 2011, customized software has been developed and employed in Grand Slam competitions. 2012 was the first year that the method was used to judge all 12 events – ten regular legs, including the finals. WDSF announced new changes on the system

based on feedback over time and the 'New System' had evolved into 'System 2.0, as the Grand Slam Series 2013 was about to start (WDSF, 2013).

Marco Sietas, former WDSF Sports Director, explained how and why they evolved the judging system. They aimed to decrease the amount of time spent on final rounds, considering the feedbacks coming from spectators, judges, and athletes. WDSF came up with a new final round solution consisting of 3 solo dances and 2 group dances which decreases the time at the final round by almost half an hour. Also, they merged PB and QM into technical qualities (TQ) criteria (Presented in Table 2.4), and thus the number of criteria was reduced to 4 from 5. In addition, the number of judges is fixed as 12 and set as three judges per criteria by randomly mapping judges to criteria components for each couple and dance so that judges only evaluate one criteria component at a time (Sietas, 2015).

Table 2. 4 TQ: Technical Qualities Criteria in Latin Dances

TQ: Technical Qualities in Latin Dances		
Brush	Hip Design	Time Step Chasse (CCC)
Foot Slip	Hip Muscular Actions	Volta Cross Chasse - version 1 (CCC)
Hold	Posture in Latin	Cucaracha in Rumba
Latin Cross	Quantity of Turn	Heel Turn in Paso Doble
Merengue Actions	Bounce in Relation to Movement	Spanish Line and Press Line (PD)
Spins and Turns in Latin	First Step of Botafogo Actions	Jive Volta Cross Chasse
Spiral Actions	Forward Walk	Flicking Action (J)
Step	Sidewalk	Cha cha cha Lock Backward (LRL)
Swivel Actions	The Bounce and Pelvic Action	Rock Actions (CCC-R)
Body Muscular Actions	Volta Actions	Forward Walk Turning (CCC-R)
Delayed Actions	Forward Walk (CCC)	Delayed Actions (CCC-R)
Foot Action	Backward Walk (CCC)	Posture in Paso Doble
Foot Placement	Sidewalk (CCC-R)	Flick Ball Change (J)

Table 2.5 demonstrates the aspects of MM, PS, and CP criteria.

Table 2. 5 Criteria for MM, PS, and CP

MM	PS	CP
Music in DanceSport	Lead through connection	Alignment
Scientific research on Timing	Shaping	Common Hip design

Timing	Couple Position (Lady to Man)	Step/Action (Number of Steps or Actions)
Bounce Timing	Lead - Hold - Shaping	Bounce in relation to the figure being danced
Rhythmic Combination in Samba		Rhythm Bounce
Samba Choreographic Timing		Forward and Backward Walk Amalgamation in Rumba
Samba Timing		Swing Jive Actions
Guapacha Timing		Jive Styles
Musical Accents in Jive		Changing the Shape of the Jive
Jive Timing		

Although the parameters are clearly set under criteria components, couples cannot see which parameter affects their evaluation process the most just by looking at the results. Also, the parameters of different criteria components are not really independent of each other because of the natural development of the dance sport (Gürbüz, 2018).

In 2015, WDSF announced that JS 2.0 had an improvement (JS 2.1) to reduce the effect of false judgments and implemented the median factor into calculations (WDSF, 2015). In JS 2.1, for every criteria component, there are three judges evaluating dancers' skills from 1 to 10, and with this new system weight of the worst and best mark is dependent on the distance to the median.

$$Distance\ Based\ Weight\ =\ 1/(1 + distance\text{^}2)\ *\ 100$$

$$Value\ of\ a\ criteria\ component\ =\ (worst * worstweight\ + median + best * bestweight)/(1 + worstweight + bestweight)$$

$$The\ total\ result\ of\ a\ dance\ =\ Value\ TQ\ +\ Value\ MM\ +\ Value\ PS\ +\ Value\ CP$$

Lastly, The JS 3.0 system is a step forward from JS 2.1. It solves the issues that come with using JS 2.1, as suggested in worlddancesport.org. In this version, 12 judges are employed again, but this time judges are divided into two groups, while six judges evaluate TQ/PS components, remaining 6 evaluate MM/CP components of criteria. The medians are selected throughout six judges' evaluations in each group, and there is a tolerance range defined based on medians, as such 1.2 for Grand Slam and Championships, 1.5 for World Open competitions. WDSF suggests that by implementing these tolerance limits, they are able to eliminate manipulative scores (WDSF, 2015).

2.4. Expert-Novice Differences in other Domains: The Case for Teaching

The differences between an expert and a novice are studied with different research methods to understand the value of experience. Dance sport is a type of sport that is evaluated by teachers. The judges of a dance sport competition consist of people who are retired dancesport performers that work as dancesport teachers. In this manner, evaluating and teaching might be

related in terms of cognitive aspects. In this section, we demonstrate the differentiation of expert and novice teachers, which is handled in various research on different settings.

Expert–novice research in combination with subject-specific context points to distinct variations in planning, instructing, and observing and reflecting on classroom occurrences; however, a comprehensive understanding of teacher cognition—namely, how experienced and beginner instructors perceive and portray educational processes such as curriculum preparation, teaching, and evaluation—remains elusive (Hogan et al., 2003).

The findings on expert-novice studies that cover teacher observations generally point out that experience has a vital role in the behaviors of teachers in a classroom setting. Hogan and Rabinowitz (2003) found that experts, in particular, were discovered to plan for the long term and to be able to make the connection between daily goals and the overall curriculum, whilst beginners seemed to concentrate on short-term planning (Hogan et al., 2003). In another study, it is found that when it comes to designing to teach a certain skill and conducting their lesson, experts were observed to be prepared more ways than novices did, and also experts preferred implementing little rehearsal before the instructional session, unlike the novice ones (Housner & Griffey, 1985). Experts were able to use a range of alternate explanations if student understanding was insufficient, whereas novices were unable to do so (Clermont et al., 1994). Experienced subject specialists were more successful than beginners in detecting classroom occurrences (Standley & Madsen, 1991). Expert teachers were interested in individual students' success when asked to comment on a lesson, while novices focused on their own instructional methods, and novices were discovered to internally plot every aspect of their class, from the questions they asked pupils to the illustrations they may use to reinforce concepts (Borko & Livingston, 1989). When compared to novices, experts made more transition between instructional methods, and were more successful in exploring student understandings, and used more supervised and observed exercise sequences to promote students' understanding (Leinhardt & Greeno, 1986). Experts concentrated on individual student accomplishments and adjusted their lessons correspondingly. Nevertheless, beginners mostly used the class's interest level as a clue to change a lesson (Housner & Griffey, 1985).

Expert and novice teachers differ on collaborating with students on a neural basis as well as a behavioral basis. (B. Sun et al., 2020). In their study, Binghai Sun and his colleagues (2020) prepared an interactive task in which they benefited from the fNIR data of participants. Participants' brain activities were recorded while they were in the task period. The results indicated that accuracy rates of experienced teacher-student dyads were higher than that of novice teacher-student when it is a cooperation task.

In our study, we used oral data of participants instead of neural data, and we used a DistilBERT model to show that participants with different levels of expertise differ significantly when speaking on a specific task.

Teachers also differ on paying attention to students that might need additional assistance regarding their experience level (Seidel et al., 2021). Teachers' reference points need to be studied to understand the cause of this differentiation (Clarridge & Berliner, 1991). The researchers used the gaze behavior of teachers to analyze and measure teachers' approaches in a classroom setting where student profiles are divided into two major groups as three incoherent (uninterested, overestimating, and underestimating) and two coherent (strong and struggling). Results show that expert teachers are much better at identifying uninterested student profiles on the first try. All of the expert teachers identify uninterested students at first try while 66.7% of novice teachers are capable of identifying uninterested student at first try.

2.5. Natural Language Processing Background & Research Areas

Natural Language Processing (NLP) is a field in which computational models and processes are run to understand human languages (Otter et al., 2021). Otter and his colleagues (2021) suggest that the NLP field can be considered as consisting of two subareas, namely core areas and applications. Core areas refer to language modeling, syntactic and semantic processing, while applications address retrieving useful information, translation and summarization of texts, and classification and clustering of documents. One of the most difficult problems in natural language processing is giving linguistic complexity to a computer to do language-based tasks correctly (Brill & Mooney, 1997). As stated in their work, Brill and Mooney (1997) mention that it is not easy for computers to distinguish the word "pen" properly between "The box is in the pen" versus "The pen is in the box (Bar-Hillel, 1964). This underlines the fact that lexical and grammatical information is not enough to understand a language; semantic, pragmatic, and world knowledge are required as well. Natural Language Toolkit (NLTK) is an open-source platform that supports semantic reasoning and brings solutions to natural language understanding problems (S. Sun et al., 2017). We benefited from NLTK while running the analysis on our datasets.

In this study, we classified Turkish and English documents by using GPT-2 Distil-BERT and Tf-Idf models. Before classifying the Turkish document, we also made a clustering in which we benefited from unsupervised machine learning to categorize the data by using the answers of participants to Likert questions.

It is much easier to find out a lemmatizer, stemmer, or a pre-trained model for natural language processing tasks when working on English texts than it is on Turkish ones. Morphemes are added to a root word in Turkish. Words are produced by affixing several suffixes to root words from a lexicon of roughly 30K root words in a highly productive manner (Oflazer, 2014). As Oflazer (2014) mentioned, there are no classes for nouns, and there are no grammatical gender marks in morphology or syntax. Also, it is very common to build up words that are equivalent to a sentence, as Oflazer (2014) stated in his review:

giz+le+y+ecek+ti+m → I was going to hide

The complex morphology and the way it interact with syntax are the main reasons for the challenges in natural language processing for Turkish. Even though these challenges remain today, recently developed natural language tools for Turkish help us get over the issues. In this study, Zeyrek lemmatizer, one of the tools that have been developed for Turkish natural language processes, has been used for lemmatization, and it will be explained with examples in the Methodology part.

Text classification can be considered as the assignment of text documents to predefined categories (Otter et al., 2021). The magnitude of the dataset and the language of the dataset take a vital role while choosing the best classifier model. Tf-Idf and Distil-BERT models were used for both of our datasets, and the results are compared to discuss which one works for them the best.

DistilBERT is a model developed by Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF (Sanh et al., 2019). It is a 60% faster, 40% smaller version of BERT with keeping 97% of the understanding skills. "BERT" stands for Bidirectional Encoder Representations from Transformers. It is a model that pre-trained on Wikipedia and Book Corpus, which is more than 800 million words (Mohd Sanad Zaki Rizvi, 2019). BERT is

working deeply bidirectional, which means it takes tokens considering their left and right side together to capture the context better.

A study shows that DistilBERT is a very useful model for argumentation retrieval (Alhamzeh et al., 2021). In this study, Alhamzeh and his colleagues (2021) aim to help users that have problems with choosing an option when given a comparative question such as “Which is better, a Mac or a PC?”. Results showed that their DistilBERT model has an accuracy of 0.8587 for the binary classification task, namely classifying arguments and non-arguments with sentences of different corpora.

In another study, DistilBERT was used to detect toxic span in text documents (Palliser-Sans & Rial-Farràs, 2021). As Sans and Farras (2021) mentioned, it is very hard to find out the aspects causing toxicity inside a text. Firstly, defining a text or phrase as toxic is a subjective act, so it can be considered as a grey area, and secondly, toxicity may not be caused by single words, but sometimes expression itself might be toxic without having any toxic word individually. BERT and Distil-BERT models were compared with their multi-depth versions. Multi-depth DistilBERT concatenates the output of different transformer blocks instead of using the last output directly. Results showed that the Distil-BERT model could be improved by adding different transformers’ outputs to some degree. For instance, in their study, Sans and Farras (2021) tried feeding the transformer blocks up to the last six transformer blocks’ output, and they had the best result by using only the last three transformer blocks’ output. They proved that the DistilBERT model can be boosted by implementing the use of multiple outputs from different transformer blocks rather than feeding the next transformer block with the output of the last transformer block.

CHAPTER 3

3. METHODOLOGY

3.1.Dataset

In this thesis study, we have two types of datasets. One is from speech data collection from experienced and non-experienced people on dancing, and the other is retrieved from reddit.com comments of a dancesport channel and lectures of former dancesport Latin branch world championship finalists on youtube.com.

3.1.1. Turkish Dataset - Expert vs. Novice

We wanted our participants to make comments on the finalist couples of the 2014 WDSF World Championship Junior 2 Latin competition held in Moscow, Russia. In addition to these comments, we wanted them to make a ranking between the couples by using the skating system of WDSF to make a quantitative analysis of the differentiation between participants and real judges. Also, we employed four components criteria of judging system 3.0 to understand how novice and experienced participants differ. These four components are technical qualities (TQ), movement to music (MM), partnering skill (PS), and choreography & presentation (CP) (Seitas, M. WDSF 2015)

We built up a model that can distinguish between expert and novice. The participants were ten experienced, licensed from Turkish Dance Sport Federation (TDSF, a member of WDSF) or retired dancesport Latin branch dancers, and 12 people that have no experience in dancing. We asked questions regarding participants' dance experiences. By asking these questions, we aimed to get information about their licenses (A is the highest, B, C, D, E is the beginner), their dancing background in dancesport, and other dances, if there are any. Also, by asking Likert questions about the level of their knowledge on criteria components (TQ, MM, PS, CP) in general and a couple of parameters inside criteria components such as rhythm interpretation, characteristics of dances, basic steps, basic figures, floor craft, and body lines & ideal form to numeric their knowledge. This enabled us to make a clustering on the data. After clustering the data into two groups based on the answers of participants to the Likert questions, we saw that the data was divided into two groups just as we pre-define novice and expert. However, the clustering process forces the data to be separated into two sections, and thus this categorization was needed to be proven statistically as well. Then, an independent t-test was done on the data, and it showed that our data does not distribute normally. Consequently, the Mann-Whitney test was applied as it is a non-parametric test. Mann-Whitney test showed that the two groups of data, which are the production of the clustering process, are significantly different, and thus we can claim the categories as expert and novice.

Figure 3.1 demonstrates data collection, pre-processing, and results comparison steps with a flowchart for the Turkish dataset. We started with data collection from 22 people including experts and novices and data collection step was followed by clustering step for categorization of the dataset. After clustering, the dataset was shaped as "balanced" (equal number of words

in each row) and “sentences” (only one sentence in each row). After the pre-processing steps, there were still typos left from lemmatization. These mistakes were removed manually. Both versions were fed into MNB and DistilBERT classifiers following the pre-processing steps. Feature extraction & vectorization, train-test data split, and classification steps are followed by order in the MNB classifier. In the DistilBERT model, fine tuning was employed with dbmdz Distilled Turkish BERT model before classification. Finally, results of DistilBERT and MNB classifiers were compared. Chapter 4 includes the reported results of these two models.

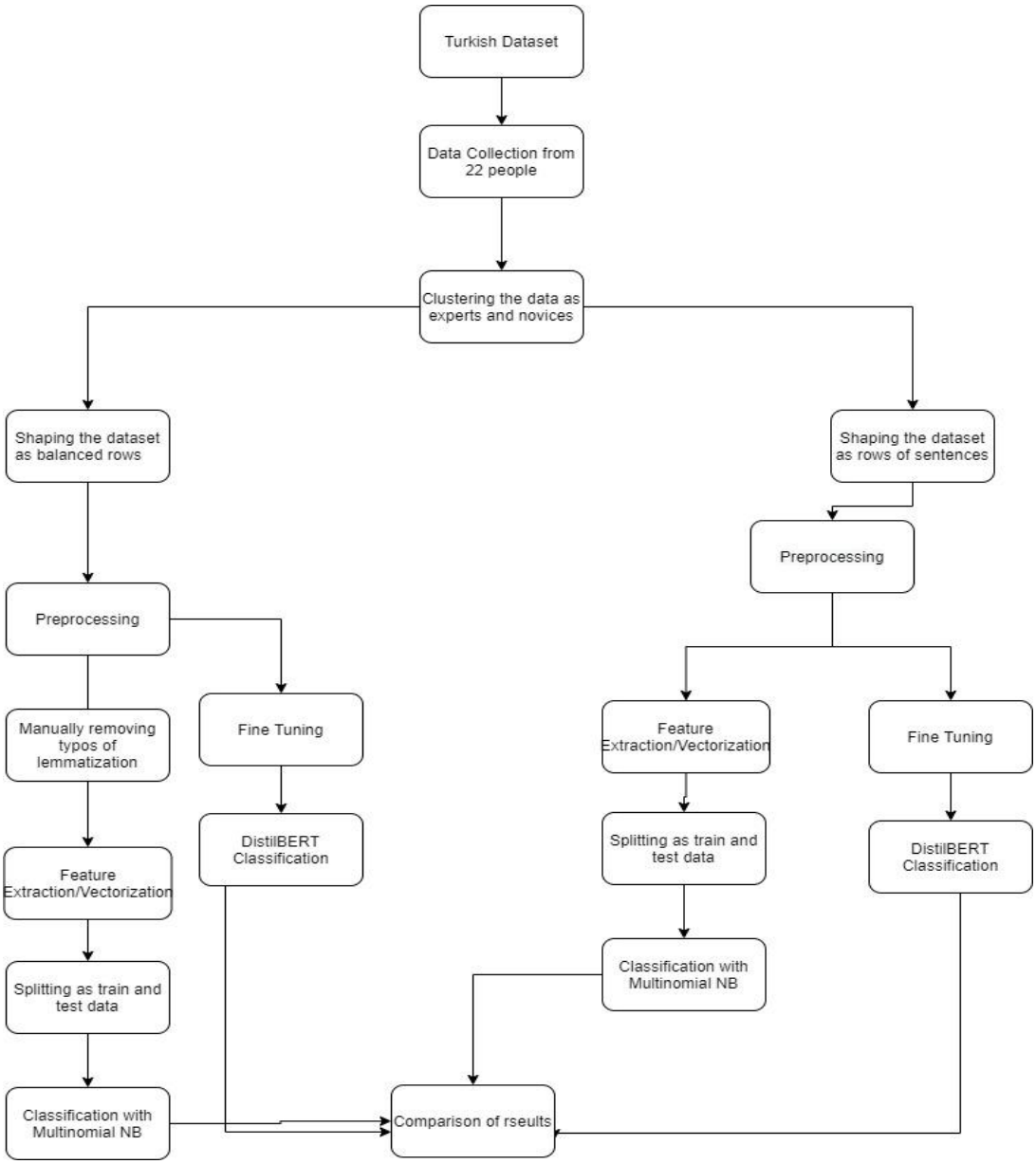

Figure 3. 1 Methodology for Turkish Dataset

3.1.2. English Dataset - Intermediate vs. Highest Degree of Expertise

The dataset coming from the web was used to see if our model also is capable of distinguishing between the intermediate level of experience, which refers to comments of people who follow dancesport channel while sharing their videos and dropping detailed feedbacks on each other’s

videos on reddit.com, and the highest level of experience referring to the lectures of top-class retired dancers of dance sport Latin branch. To build up an equally distributed data set, which is named as “Balanced Dataset” in this study, on both intermediate and highest level of experience, comments on reddit.com and transcript of lecturers on youtube.com have been divided into lines and fed to an Excel document as such 149 lines of comments and 150 lines of the transcript.

Figure 3.2 demonstrates the methodology for the English dataset. The lemmatizer used for the English dataset, WordnetLemmatizer, did not require a further cleaning on the data. Thus, right after the pre-processing, the datasets were fed to feature extraction and fine-tuning steps for MNB and DistilBERT classifiers respectively.

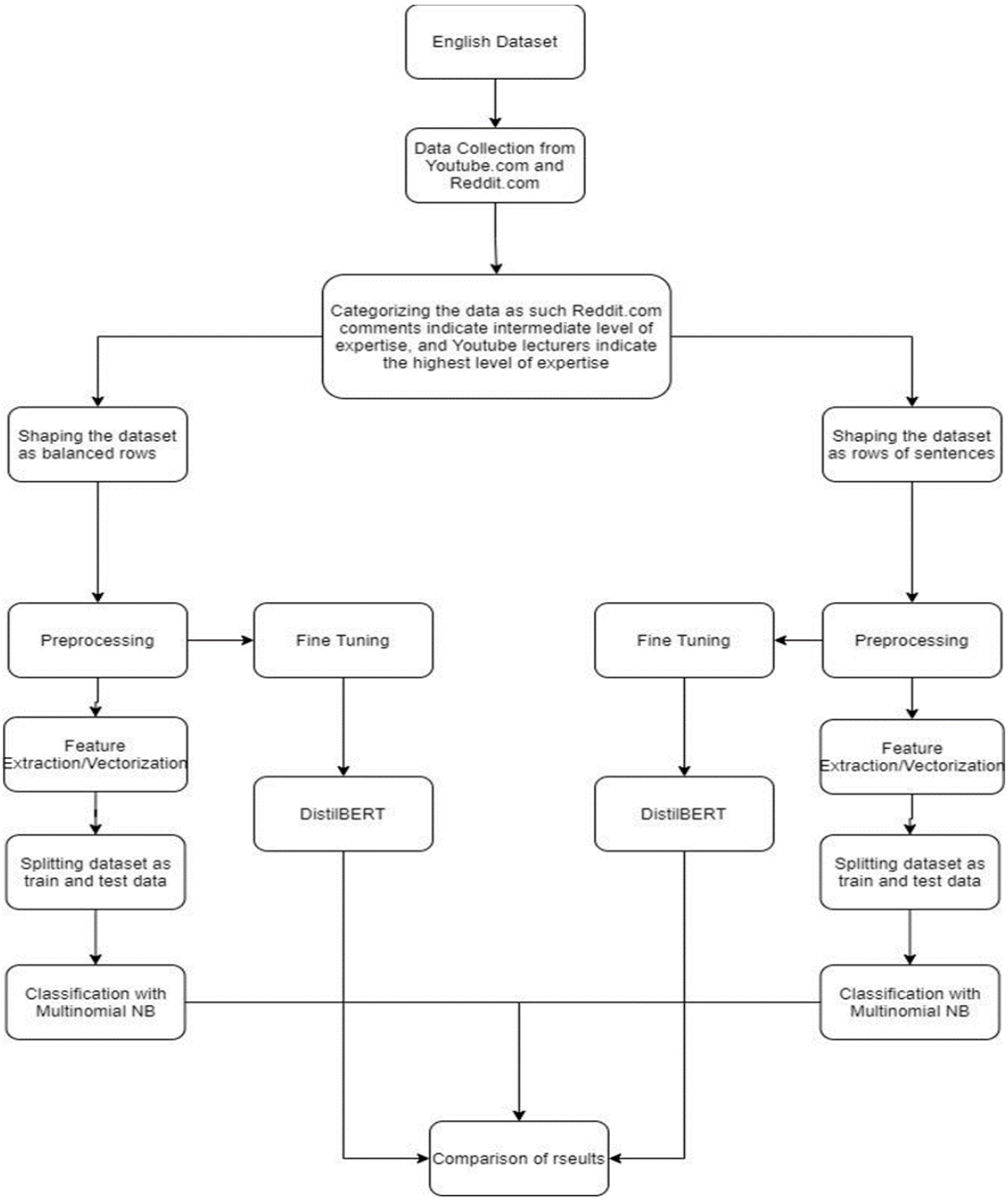

Figure 3. 2 Methodology for English Dataset

3.2. Preprocessing

Before training and testing our model, data should be cleaned to have accurate results. Preprocessing steps include lowering letters, sentence tokenizing, removing stop words, and lemmatization of the words.

3.2.1. Lowering Letters & Cleaning Syntax

The first step of the preprocessing is lowering the letters and removing unwanted characters. Since there might be characters that are accidentally put into the data set while converting

speech data to text data, we need to make sure that these characters should be removed. Also, to have a standard within the dataset and get over the case-sensitive processes, if there are any, we lowered the letters. Text lowering and cleaning unwanted characters processes are independent of the language for our study. Both for the Turkish and English datasets, we applied the same function. Unwanted characters include [_"\-;%()|+&=*%.,!?:#$@\[\]/:\/\/.*[\r\n]*

3.2.2. Removing Stopwords

According to a web article (Stopwords in SearchWorks - to Be or Not to Be, 2011), stopwords are very common words in a language that do not help us classify/distinguish within a dataset. Stopwords can be imported as a package and run on a dataset. Although stopword packages cover lots of stopwords as a list, this list can be extended for different circumstances. In our study, we used two different stopwords lists for Turkish and English datasets. Also, we extended the English stopwords list since one of the two categories of our English speech data set includes numbers.

Turkish stopwords list that is covered in NLTK library of Python is as follows (www.nltk.org);

['acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey', 'biz', 'bu', 'çok', 'çünkü', 'da', 'daha', 'de', 'defa', 'diye', 'eğer', 'en', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'için', 'ile', 'ise', 'kez', 'ki', 'kim', 'mı', 'mu', 'mü', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye', 'niçin', 'niye', 'o', 'sanki', 'şey', 'siz', 'şu', 'tüm', 've', 'veya', 'ya', 'yani']

English stopwords list that has been used in the study is as follows (NLTK, 2021)

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", 'the','one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight']

English data set consists of two groups of people. One group is commentators of videos of a dance sport athlete on reddit.com, and this group is considered as "intermediate level expertise," the other group is two retired dancesport champions (currently judges and teachers) that are giving a lecture by interpreting the common mistakes that couples do on the dance floor and this group considered as "highest level of expertise." In these lectures, teachers want the class makes practice occasionally. In these practices, teachers start to count 'one' 'two' 'three', 'four', 'five', 'six', 'seven', 'eight' as the bar consists of 8 numbers music-wise. We try to see if our model is capable of distinguishing intermediate and highest level of expertise, and thus we removed the parts that the class is practicing as it is not related to experience, and it was not relevant to the context.

3.2.3 Lemmatization

Lemmatization and stemming are used for reaching out to the origin and the root of the word. While stemming mainly removes prefixes and suffixes of the word, lemmatization looks for the origin of the word semantically.

In this study, we used WordNetLemmatizer (NLTK, 2021) from the NLTK library for the English dataset and Zeyrek Lemmatizer (Bulat, 2020) for the Turkish dataset.

Before we ran the Tf-Idf model, we cleaned and lemmatized the data. Since the lemmatizer we used, Zeyrek, gives an output of all possible roots of a word as a list, we selected the useful words within the lemmatized words list. Also, the typos and wrong lemmatization were fixed manually after preprocessing. Then, we fed the cleaned data to the model. Table 3.1 presents the Turkish dataset before and after lemmatization.

Table 3. 1 Raw text vs. Lemmatized text for Turkish Dataset

Turkish Text	Lemmatized
Şu anda vücut porsiyonları yani üst gövde göre bacak oranı çok fazla, uzun göstermek çalışmışlar.	[('Şu', ['şu']), ('anda', ['an', 'ant']), ('vücut', ['vücut']), ('porsiyonları', ['porsiyon']), ('yani', ['yani', 'Yani']), ('üst', ['üst']), ('gövdeye', ['gövde']), ('göre', ['görmek', 'göre']), ('bacak', ['bacak']), ('oranı', ['ora', 'oran']), ('çok', ['çok']), ('fazla', ['faz', 'fazla']), (',', [',']), ('uzun', ['uz', 'uzun', 'Uz']), ('göstermeye', ['göstermek']), ('çalışmışlar', ['çalışmak']), ('.', ['.'])]
Erkeğin bacak hareketleri çok daha hızlı gibi şu an için.	[('Erkeğin', ['erkek']), ('bacak', ['bacak']), ('hareketleri', ['hareket']), ('çok', ['çok']), ('daha', ['daha']), ('hızlı', ['hız', 'hızlı']), ('gibi', ['gibi']), ('şu', ['şu']), ('an', ['an', 'anmak']), ('için', ['iç', 'içmek', 'için']), ('.', ['.'])]
Kadın daha kol ve üst gövdeyle dans ediyor gibi.	[('Kadın', ['kadın', 'kadı']), ('daha', ['daha']), ('kol', ['kol']), ('ve', ['ve']), ('üst', ['üst']), ('gövdeyle', ['gövde']), ('dans', ['dans']), ('ediyor', ['etmek']), ('gibi', ['gibi']), ('.', ['.'])]
Özellikle oğlanın alt tarafta seri olarak hareketleri devam ediyor.	[('Özellikle', ['özellikle']), ('oğlanın', ['oğlan']), ('alt', ['alt']), ('tarafta', ['taraf']), ('seri', ['seri', 'ser', 'Seri']), ('olarak', ['olmak']), ('hareketleri', ['hareket']), ('devam', ['devam', 'deva', 'Deva']), ('ediyor', ['etmek']), ('.', ['.'])]
Bunlar beraber yürüdüler, diğerinde biraz daha kopukluk vardı.	[('Bunlar', ['bun', 'bu']), ('beraber', ['beraber']), ('yürüdüler', ['yürümek']), (',', [',']), ('diğerinde', ['diğer']), ('biraz', ['biraz']), ('daha', ['daha']), ('kopukluk', ['kopuk']), ('vardı', ['varmak', 'var']), ('.', ['.'])]

We noticed the pattern that points out the desired lemmatization output within the Turkish lemmatized list of words. As it can be referred in Table 3.1, the first element of a lemmatized word is the word itself, the second argument is a list of possible origins of the word.

The first row of Table 3.1 has the text “Şu anda vücut porsiyonları yani üst gövde göre bacak oranı çok fazla, uzun göstermek çalışmışlar.”. To have an illustration, the word “anda” can be investigated.

“Şu anda” phrase refers to “at the moment” in English. “An” refers to “moment” and “-da” suffix may refer to “at”. Lemmatizer claims two possible origins for the word “anda” here and gives an output of the word itself and a list of possible origins.

Lemmatizer output for “anda” : ('anda', ['an', 'ant'])

This is a list consists of two arguments as such the first argument is an element (the word itself), and the second argument is a list of possible origins of that word. In this example we have ‘an’ and ‘ant’ as possible word origins. The lemmatizer gives us two possibilities. Either the word origin was ‘ant’ (oath) and it took the suffix ‘-a’, or the word origin is ‘an’ (moment) and it took the suffix ‘-da’. In this context, the desired word origin is ‘an’ and it refers to the first element of the second argument in the output of lemmatizer. We noticed that almost all desired origins refer to the first element of the second argument. Thus, we have selected those ones by implementing the code below.

```
df["selected"] = df["analyzed"].apply(lambda x: [x[1][0] for x in x])
```

After applying the code line above, we were able to append desired word origins. Nevertheless, there were still some outputs that do not represent the origin of the word. We had to manually clean those ones. Table 3.2 presents an example of the Turkish dataset after lemmatization and manual cleaning of the data with word-by-word translation to English.

Table 3. 2 Comparison of raw text and preprocessed text for Turkish data set

Raw Text	**Preprocessed Text**
Şu anda vücut porsiyonları yani üst gövdeye göre bacak oranı çok fazla, uzun göstermeye çalışmışlar. Translation: At the moment, their body parts, that is, the ratio of legs to upper body are too high, they tried to make them look longer.	an vücut porsiyon üst gövde göre bacak oran fazla uzun göstermek çalışmak Translation: moment body parts the ratio legs upper body too high, try to make look long.
Erkeğin bacak hareketleri çok daha hızlı gibi şu an için Translation: The male's leg movements seem much faster for now	erkek bacak hareket hız an Translation: male leg movement fast now
Kadın daha kol ve üst gövdeyle dans ediyor gibi. Translation: The woman is more like dancing with her arms and upper body.	kadın kol üst gövde dans etmek Translation: woman arm up body dancing
Özellikle oğlanın alt tarafta seri olarak hareketleri devam ediyor Translation: Especially the boy's movements continue in series at the bottom.	Özellik oğlan alt taraf seri olarak hareket devam etmek Translation: Speciality boy bottom movement continuing series bottom.
Bunlar beraber yürüdüler, diğerinde biraz daha kopukluk vardı. Translation: These walked together, with the other a little more disconnected.	beraber yürümek diğer biraz kopuk var Translation: walk together, other a little disconnect

Bunlar seyirciye daha dönüktü tüm seyircilere bakmaya çalıştılar. Translation: They were more focused on the audience; they tried to look at all the audience.	seyirci dönük seyirci bakmak çalışmak Translation: focused audience try to look at audience
Bunun vücut oranında üst gövdeyi uzatmaya çalışmışlar gibi geldi. Translation: It felt like they were trying to lengthen the upper body at the body rate of this.	vücut oran üst gövde uzatmak çalışmak Translation: body ratio upper body lengthen try

For the English dataset, we fed the system by following the same methods we have done for the Turkish dataset. However, as it can be seen on Table 3.3, this time the step for selection of origin words was not needed as the English lemmatizer gives only one output (lemmatized word) for every input.

Table 3. 3 Raw texts vs. Lemmatized text for English Dataset.

English Text	Lemmatized
Hi, stop being so negative about yourself. Your dancing is not bad. I think you have good technique for the level of comp you are in. Ok, first easiest thing. You look like you are	['Hi', 'stop', 'be', 'so', 'negative', 'about', 'yourself', 'Your', 'dancing', 'be', 'not', 'bad', 'I', 'think', 'you', 'have', 'good', 'technique', 'for', 'the', 'level', 'of', 'comp', 'you', 'be', 'in', 'Ok', 'first', 'easy', 'thing', 'You', 'look', 'like', 'you', 'be']
dancing solo, and there just happens to be someone dancing around you. Rumba is meant to be the dance of love, but yours is devoid of emotion. In a competition you are performing	['dance', 'solo', 'and', 'there', 'just', 'happen', 'to', 'be', 'someone', 'dance', 'around', 'you', 'Rumba', 'be', 'mean', 'to', 'be', 'the', 'dance', 'of', 'love', 'but', 'yours', 'be', 'devoid', 'of', 'emotion', 'In', 'a', 'competition', 'you', 'be', 'perform']
you shouldn't be thinking about all the technique you need to do. If you don't look like you are having fun, no-one watching you will have fun either, they will feel your	['you', 'shouldn', 't', 'be', 'think', 'about', 'all', 'the', 'technique', 'you', 'need', 'to', 'do', 'If', 'you', 'don', 't', 'look', 'like', 'you', 'be', 'have', 'fun', 'no', 'one', 'watching', 'you', 'will', 'have', 'fun', 'either', 'they', 'will', 'feel', 'your']
discomfort. Further I'd look at good dancers doing the different dances and listen to lots of music. Work out what you think makes each dance different then do it in yours, because	['discomfort', 'Further', 'I', 'd', 'look', 'at', 'good', 'dancer', 'do', 'the', 'different', 'dance', 'and', 'listen', 'to', 'lot', 'of', 'music', 'Work', 'out', 'what', 'you', 'think', 'make', 'each', 'dance', 'different', 'then', 'do', 'it', 'in', 'yours', 'because']
at the moment, bar the steps being different I don't see a difference. I think this is likely due to an overfocus on technique. Techniquewise, I think that you are using your moving	['at', 'the', 'moment', 'bar', 'the', 'step', 'be', 'different', 'I', 'don', 't', 'see', 'a', 'difference', 'I', 'think', 'this', 'be', 'likely', 'due', 'to', 'an', 'overfocus', 'on', 'technique', 'Techniquewise', 'I', 'think', 'that', 'you', 'be', 'use', 'your', 'move']

Table 3.4 demonstrates the pre-processed texts to the corresponding raw texts for the English dataset. It should be noted that after removing stopwords the data noticeably got smaller.

Table 3. 4 Comparison of raw text and preprocessed text for English dataset

Raw text	**Preprocessed text**
you look like you are dancing solo, and there just happens to be someone dancing around you.	look like dance solo happen someone dance around
rumba is meant to be the dance of love, but yours is devoid of emotion.	rumba mean dance love devoid emotion
in a competition you are performing you shouldn't be thinking about all the technique you need to do.	competition perform thinking technique need
if you don't look like you are having fun, no-one watching you will have fun either, they will feel your discomfort.	look like fun watch fun either feel discomfort
work out what you think makes each dance different then do it in yours, because at the moment, bar the steps being different i don't see a difference.	work think make dance different moment bar step different see difference
technique wise, i think that you are using your moving leg and not the standing leg to create movement.	technique wise think use move leg stand leg create movement
this makes your dancing look too fast and out of control (you can use more time on the action of each step).	make dance look fast control use time action step

3.3. Models

We have used Tf-Idf and DistilBERT models to find out the best model that helps us distinguish experts from less experienced for both the English and the Turkish datasets. Since we have limited amount of data, we put our data into models in various ways.

3.3.1. Multinomial Naïve Bayes (Tf-Idf)

This method employs the naïve Bayes algorithm for multinomially distributed data. The data are represented as word vector counts in this method, but Tf-Idf vectors can also be employed (Pedregosa et al., 2011).

In the Turkish dataset, we have 22 participants (10 novices, 12 experts). When we put this dataset into the model without splitting, it makes only 22 rows of data which we cannot run proper testing. In order to increase the number of inputs we can test, first, we divided our data into sentences and then fed to the model.

3.3.2. Feature Selection

We used TfidfVectorizer from Sci-Kit Learn, and we had a couple of trials on feature selection methods. First, we started with `SelectPercentile` (SP) from Sci-Kit Learn. To illustrate, when SP is set to 10, it enables your model to focus on only the most successful 10% of features that are good at indicating the categories and ignore other features. Secondly, we had a couple of trials on setting a `max_features` value. `max_features` generates a vocabulary that only considers the top `max_features` in the corpus that are ranked by term frequency. Finally, we used `max_df` to select features. If `max_df` is set to 0.2, it checks the words and ignores the ones that exist in 20% and more documents. The logic behind using `max_df` is very similar to that of the "stopwords" function. There might be some unrelated words that are not stopwords but still very common in documents and not very useful for category selection. `max_df` enables the model to ignore those kinds of words.

We split the data as test size is 0.2 and ngram range of (1,2). Then we used Multinomial Naive Bayes from (MNB) Sci-Kit Learn to make the classification. We took the outputs of precision, recall, and accuracy scores for both categories. Table 3.5 presents the shapes of the English and the Turkish datasets as train and test sizes.

Table 3. 5 Shapes of English and Turkish datasets both as balanced rows and sentences for Tf-Idf

	Balanced Rows	1 Sentence/Row
Turkish Dataset Train	124	604
Expert(E)-Novice(N) Distribution	64 E, 60 N	307 E, 297 N
Turkish Dataset Test	31	151
Expert(E)-Novice(N) Distribution	15 E, 16 N	75 E, 76 N
English Dataset Train	239	392
Expert(E)-Novice(N) Distribution	119 E, 120 N	198 E, 194 N
English Dataset Test	60	98
Expert(E)-Novice(N) Distribution	31 E, 29 N	57 E, 41 N

The supervised learning algorithms based on Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable are known as naive Bayes methods. MNB applies the naive Bayes method for multinomially distributed data (scikit-learn, n.d.). Table 3.6 demonstrates the accuracy scores of the MNB classifier for the Turkish datasets. Chapter four can be referred to see precision and recall values for the trials came up with the highest accuracy score.

Table 3. 6 Accuracy Scores of Turkish Datasets across different feature selection criteria

Feature Selection Methods	Number of Features Selected - Turkish Dataset Balanced	Accuracy Score of Turkish Dataset Balanced	Number of Features Selected - Turkish Dataset Sentences	Accuracy Score of Turkish Dataset Sentences
SP = 5	167	0.84	177	0.68
SP = 10	334	0.87	353	0.70
SP= 20	667	0.90	704	0.75
max_df = 0.1	3292	0.84	3513	0.79
max_df = 0.3	3327	0.90	3522	0.77
max_df = 0.5	3336	0.90	3522	0.77
max_features = 100	100	0.94	100	0.65
max_features = 500	500	0.90	500	0.74
max_features = 1000	1000	0.90	1000	0.73

Table 3.7 presents the accuracy score of the English Dataset with all feature selection configurations. Configurations that have the highest accuracy scores were presented in Chapter four. Remaining results are reported in Appendix C.

Table 3. 7 Accuracy Scores of English Datasets across different feature selection criteria

Feature Selection Methods	Number of Features Selected - English Dataset Balanced	Accuracy Score of English Dataset Balanced	Number of Features Selected - English Dataset Sentences	Accuracy Score of English Dataset Sentences
SP = 5	330	0.92	296	0.87
SP = 10	660	0.88	591	0.83
SP= 20	1320	0.82	1180	0.84
max_df = 0.1	6542	0.88	5881	0.86
max_df = 0.3	6589	0.82	5900	0.82
max_df = 0.5	6596	0.85	5904	0.81
max_features = 300	300	0.93	300	0.85
max_features = 500	500	0.95	500	0.89
max_features = 1000	1000	0.95	1000	0.86

These three feature selection parameters were used and manipulated separately since `SelectPercentile` dominates `max_features` and `max_df`. Also, the features coming from `max_features` and `max_df` can be contradictory since one of them tries to ignore the most common words while the other vectorizes the most seen words in the document.

Before running the MNB model, we lowered the letters, removed stopwords and punctuations, lemmatized the words and appended them to have a clean list of words. For the Turkish dataset, we manually fixed the data after the lemmatization step; however, manual data fixing was not needed for the English dataset.

3.3.3. DistilBERT Model

DistilBERT is a distilled version of BERT, namely Bidirectional Encoder Representations from Transformers. We used dbmdz Distilled Turkish BERT model for our Turkish dataset. This model was trained on 7GB of the training data that was used to train DistilBERTurk, which is a cased distilled model-driven for Turkish. DistilBERTurk was trained with Hugging Face for 5days (Hugging Face, 2021)

We used dbmdz/bert-base-cased-finetuned-conll03-English for our English dataset. It is a commonly used model for English NLP studies.

We have split our data as train, test, and validation. The validation data are not involved training and test sessions. The accuracy output comes regarding how accurate our model is on matching the test data to the correct categories after the training. The validation data can be considered as a real-life example for us to see if our model is accurate or not. The size of the validation data is not really important since it doesn't affect the accuracy value. However, setting the percentile of validation data too big may cause lowering the test data, which results in lowering the sensitivity of the model. We fed the model with two different shapes of Turkish data, just as we did in the MNB Tf-Idf model. First, we fed the model with the data that was distributed equally across rows. Secondly, we fed the model with the same dataset but sentences across rows. Since we do not have too many rows of data, especially in the balanced version, we only split the dataset into two as such test and train data. We used validation data for the datasets that were distributed across rows as sentences. Table 3.8 presents the shapes of the datasets fed to the DistilBERT model. English dataset is always fed to the model as one sentence per row.

We set the batch size as 16 and the number of epochs as 3. We kept the epoch value low since increasing the epoch leads to overfitting.

Table 3. 8 Shapes of datasets for DistilBERT Model

	Turkish Dataset Balanced	Turkish Dataset Sentences	English Dataset Sentences
Train	120	631	344
Test	35	100	120
Validation	-	25	25

3.3.4. Increasing the Test Dataset Size

The data that is Turkish were collected from 22 people, and this means 22 rows of data in total. When we wanted to split our data as test and train, we noticed that it was impossible to have a sensitive model if we followed the common train-test split threshold, which is 0.2-0.3 for the test data. To overcome this challenge, we divided our data into two different ways.

First, we divided the data into rows in a way that each row had almost the same number of words. 155 rows of data were arranged to be fed to the model in the Turkish dataset, and 299 rows of data were fed to the model in the English dataset.

Secondly, we split the sentences in the datasets and fed the datasets as each row has one sentence. We had 756 rows of data in the Turkish dataset, while we had 489 rows of data in the English dataset.

3.4. Methodology Summary

In this section, we explained the steps we followed through data collection, pre-processing, feature selection, and data shaping. For the Turkish dataset, we collected speech data from 22 participants including ten experts and 12 novices. The data was converted to text data and preprocessed by lowering the letters, removing stopwords and punctuations, lemmatizing the words, and fixing the errors left from lemmatization manually. The manuel fixing of the English dataset was not needed since WordNetLemmatizer gave desired outputs for lemmatization step.

After the preprocessing steps, the datasets were fed to the models in different shapes. By manipulating `SelectPercentile, max_df, and max_features` parameters, various trials were employed in MNB model and the highest accuracy scores are reported in Chapter 4. For the DistilBERT model, `learning rate` parameter was manipulated throughout the trials.

CHAPTER 4

RESULTS

In this chapter, we demonstrate the outcomes of Tf-Idf and DistilBERT classifiers. These classifiers were fed with the Turkish dataset we have collected from 22 people consisting of novice and experts on dancing and the English dataset that was collected from reddit.com and youtube.com. The same Tf-Idf model was used for both Turkish and English data. Different DistilBERT models were used for Turkish and English datasets separately.

4.1. Setup

The data collection process was done via an online meeting recording on Microsoft Teams for the Turkish dataset. People watched a video of the final Samba presentation of the World Championship Junior-2 Latin competition, which was held on 30 March 2014, Moscow (InterDance.Ru, 2014). The recordings of participants were converted to text data and fed to the model after preprocessing steps.

English dataset was collected from comments to a dancesport performer's video (Critique for Gold Am/Am Latin, 2018) and a lecture of a dancesport judge/teacher/former champion (Slavik Kryklyvyy On The Importance Of Musicality and Rhythm, 2021).

The Turkish dataset was clustered based on the answers to Likert questions regarding dance sport experience. The results of clustering are reported in Table 4.1.

Table 4. 1 K-Means Clustering

Cluster	Experts	Novices
Size (the number of participants)	10	12
Explained proportion within cluster heterogeneity	0.333	0.667
Within sum of squares	7.220	14.457
Silhouette score	0.547	0.393
Centroid Participant ID	-0.108	0.090
Centroid Total Likert	1.047	-0.873

The clustering procedure divides the data into two sections, this categorization had to be statistically validated as well. After we ran the test of normality, we saw that our data were not

distributed normally. A Shapiro-Wilk test showed a significant departure from normality for cluster Experts, $W(10)=0.809$, $p=0.012$.

Since significant results suggest a deviation from normality, we ran a Mann-Whitney test, and Mann Whitney test indicated that the difference between two clusters is statistically significant, $U(N_{novice}=12, N_{expert}=10)=0$, $p<0.01$. Thus, categorization can be done in light of this clustering process.

4.2. Tf-Idf, Multinomial naïve Bayes Classifier

While running the Tf-Idf classifier, `max_df` and `SelectPercentile` values were manipulated in combination, whereas `max_features` was used and manipulated separately. In this section, we demonstrate how precision, recall, and accuracy scores change due to the changes made on `max_df` , `SelectPercentile`, and `max_features`.

The results are shared in tables with corresponding `max_df`, `SelectPercentile`, and `max_features` parameters.

Precision=True Positive /Total Predicted Positive

Recall = True Positive /Total Actual Positive

Accuracy-Score = 2*(Precision*Recall)/(Precision+Recall)

4.2.1. Results from the English Data

Only one table has been shared within three configurations of each parameter. Table 4.2 demonstrates the best accuracy,precision, and recall results among the versions in which `SelectPercentile` was set to five.

Table 4. 2 English Dataset MNB, SelectPercentile =5

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.96	0.95	0.87	0.79	0.91	0.86	30	52
Category 1	0.88	0.80	0.97	0.96	0.92	0.87	30	46
accuracy					0.92	0.87	60	98
Macro avg.	0.92	0.88	0.92	0.87	0.92	0.87	60	98
Weighted avg.	0.92	0.88	0.92	0.87	0.92	0.87	60	98

When `SelectPercentile` parameter was used and manipulated, the best accuracy scores for the English dataset were 0.92 for the balanced dataset where `SelectPercentile` value is 5. This means our model classifies at its best when it works with only the best 5% of all

features. Also, when `SelectPercentile` is manipulated, our balanced dataset is more suitable than the other dataset in which we placed 1 sentence to each row.

As it can be seen on Table 4.3, when `max_df` parameter is used and manipulated, the best accuracy scores for the English dataset is 0.88 for the balanced dataset where `max_df` value is 0.1. This means our model classifies at its best when it ignores the words that exist in more than 10% of documents. Also, when `max_df` is manipulated, our balanced dataset is again more suitable than the dataset in which we placed one sentence per row.

Table 4. 3 English Dataset MNB, max_df=0.1

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.96	1.00	0.80	0.73	0.87	0.84	30	52
Category 1	0.83	0.77	0.96	1.00	0.89	0.87	30	46
accuracy					0.88	0.86	60	98
Macro avg.	0.89	0.88	0.88	0.87	0.88	0.86	60	98
Weighted avg.	0.89	0.89	0.88	0.86	0.88	0.86	60	98

Table 4.4 presents the best accuracy scores (0.95) came out from max_feature trials when it was set to 1000 with balanced English dataset.

Table 4. 4 English Dataset MNB, max_features =1000

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.80	0.96	0.88	0.95	0.84	27	41
Category 1	0.97	0.91	0.94	0.84	0.95	0.87	33	57
accuracy					0.95	0.86	60	89
Macro avg.	0.95	0.85	0.95	0.86	0.95	0.85	60	98
Weighted avg.	0.95	0.86	0.95	0.86	0.95	0.86	60	98

The results showed that our model is pretty good at classifying two different groups across the dataset, which are coming from video comments of people who are interested in dancesport and from video transcripts of former champion present dancesport adjudicators. Among all

trials, `max_features` was the parameter that gave the highest accuracy when it was set to 1000. For the remaining results, Appendix C can be referred to.

4.2.2. Results from the Turkish Data

The results of the Turkish Dataset when `SelectPercentile` was set to ten is reported in Table 4.5. We noticed that precision and recall values were equal to 1.00 when `SelectPercentile` was set to 5 and 20. By changing the random-state value we aimed to get away from the risk of overfitting.

Table 4. 5 Turkish Dataset MNB, SelectPercentile =10

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.69	0.69	0.92	0.70	0.79	0.69	12	73
Category 1	0.93	0.71	0.74	0.71	0.82	0.71	19	78
accuracy					0.81	0.70	31	151
Macro avg.	0.81	0.70	0.83	0.70	0.80	0.70	31	151
Weighted avg.	0.84	0.70	0.81	0.70	0.81	0.70	31	151

After the trials with different `SelectPercentile` values, `max_df` parameter was selected to be configured. We started to the `max_df` trials by setting it to 0.1. Then 0.3 and 0.5 values were set to find out the most accurate result. The accuracy scores when `max_df` was set to 0.1 and 0.3 were 0.94. When `max_df` is set to 0.1, it operates with only the features that do not exist in more than 10% of the documents. Results of `max_df`=0.1 is reported in Table 4.6, for the results of `max_df`=0.3 and `max_df`=0.5 Appendix C can be referred to.

Table 4. 6 Turkish Dataset MNB, max_df=0.1

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.80	0.93	0.77	0.93	0.78	14	73
Category 1	0.94	0.79	0.94	0.82	0.94	0.81	17	78
accuracy					0.94	0.79	31	151
Macro avg.	0.93	0.80	0.93	0.79	0.93	0.79	31	151
Weighted avg.	0.94	0.79	0.94	0.79	0.94	0.79	31	151

Manipulating the `max_df` value between 0.1 and 0.5 did not affect the results for balanced and Turkish dataset. Also, there was no difference in the outcomes of Turkish sentence dataset when `max_df` was set to 0.3 and `max_df` was set to 0.5. `Max_df` parameter is used to eliminate the words that are existing more than certain percentile of the documents. This indicates that there are not many words existing more than 10% of the documents.

Table 4. 7 Turkish Dataset MNB, max_features=1000

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.73	0.88	0.68	0.90	0.71	16	73
Category 1	0.88	0.73	0.93	0.78	0.90	0.74	15	78
accuracy					0.90	0.73	31	151
Macro avg.	0.90	0.73	0.90	0.73	0.90	0.73	31	151
Weighted avg.	0.91	0.73	0.90	0.73	0.90	0.73	31	151

Among all trials that were run on MNB classifier model by manipulating SP, max_df and max-features parameters separately, the balanced dataset with max_df set to 0.1 and 0.3 ended up with the highest accuracy scores (0.94) for the Turkish dataset.

Balanced datasets were adjusted in a way that each row has almost equal length of word series. Both for Turkish and English datasets, balanced ones were more likely to end up with higher accuracy scores in our MNB classifier.

4.3. DistilBERT Model Results

While running trials on DistilBERT model, learning rate parameter was manipulated and corresponding accuracy & entropy loss results were recorded.

Learning rate is the amount that DistilBERT model updates its weights after using examples of training dataset. If learning rate is set too small, we need lots of iterations to find out the best value. If learning is set too large, then the possibility of overshooting the best value increases. Thus, setting a proper learning rate is important.

Starting from 5e-6, we increased learning rate in two more steps to 5e-4. Table 4.8 demonstrates the highest accuracy which was achieved when learning rate was set to 5e-5.

Table 4. 8 DistilBERT Model results when learning rate was set to 5e-5.

	Turkish Dataset Balanced	Turkish Dataset Sentences	English Dataset Sentences
Accuracy-Score	0.91	0.72	0.76
Cross-Entropy Loss	0.19	0.70	0.54

The highest accuracy-score for Turkish dataset was 0.91 with a loss value of 0.19 and it came out where learning rate is set to 5e-5 for the balanced data. For the English dataset, 5e-6 led to the highest accuracy which was 0.85 with a loss value of 0.35 where balanced data was used.

Cross-entropy loss indicates how much our model diverges from the actual category while predicting. To have an illustration, if the actual category of the test data is 0 and model gives 0.85, then there would be a high cross-entropy loss. In our model we used SparseCategoricalCrossentropy parameter from keras with log-loss argument.

4.4. Participants' Ranking Results

Turkish dataset was collected from 22 people, and it was clustered to have a reasonable categorization. After clustering we had 12 novices and 10 experts. We wanted all participants to rank six couples that are dancing Samba in the final round of World Championship Junior-2 Latin competition, which is held on 30 March 2014, Moscow (InterDance.Ru, 2014). Ranking data of participants were collected considering their categories. This data was used to create a product ranking of each category and make a comparison between the actual ranking and these rankings. While creating group rankings COUNTIF and SUMPRODUCT tools were used in Microsoft Excel. The results of the rankings made by novices and experts are reported in Table 4.9 with Actual Ranking (The ranking report that was filled by the adjudicators in World Championship Junior-2 Latin competition 30 March 2014).

Table 4. 9 Couples' Ranking by Participants vs. Adjudicators

Couple Number	Actual Ranking	Novice Ranking	Expert Ranking
15	6	2	3
37	1	4	1
44	5	1	4
51	2	5	2
66	3	3	6
79	4	6	5

While expert group estimates the 1st and 2nd couples' positions correctly, novice group could only estimate 3rd place. Experts' rankings were more compatible when compared to that of novices.

4.5.Evaluation of the Results

In our thesis, we collected Turkish speech data from 22 people including experts and novices on dance sport and collected English text and transcript data from Reddit.com and Youtube.com respectively. We used MNB classifier and DistilBERT classifier by making changes on different parameters such as "SelectPercentile", "max_df", "max_features" and "learning rate".

Since we have limited amount of data, we shaped our dataset in a way that we can increase the number of test documents to be able to achieve more sensitive classifier. To do that, we fed the model in two different ways. First, we divided the data into the rows of equal number of words, and we called it "balanced dataset". Secondly, we split the dataset into sentences which resulted in increasing the number of rows even more than the balanced dataset. We called the latter as "sentences dataset". To compare the balanced and sentence-shaped datasets, an independent-samples t-test was conducted. The accuracy scores of balanced datasets (M=0.88, SD=0.042) was significantly higher than the sentences dataset (M=0.789, SD=0.070), t(18)=5.143, p<0.001.

Figure 4.1 draws that balanced version have better results on both in Tf-Idf and DistilBERT models for both Turkish and English datasets.

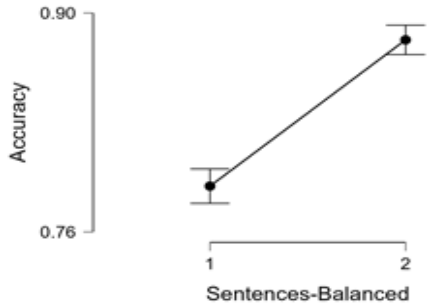

Figure 4. 1 Sentence and Balanced Data vs. Accuracy

The focus of expert and novice people were on different aspects while making comments to the video demonstrated. In Turkish dataset, experts made critical comments on dancers' joints such as ankles, knee, elbows, and wrists. However, novice group did not go deeper than limbs while watching the video. It is probably because experts know that joints are the key parts of

the body while creating a proper move and in case of a failure in dancing, they try to address the problem to the joints as a part of root cause analysis.

When there is a sudden quick move and a sudden stop in couples' choreography, novice people generally prefer describing the couple as "hızlı" (fast), however, experts tend to use different terms for this kind of moves, namely "net" (clear) and "keskin" (sharp). These terms refer to both starting and ending a move very fast as well as they refer to neat and tidy dancing.

Experts declare that they focus more on their fellows on the couples. It might be related to the dance education and experience they had in the past. Male dancers and female dancers do the same basic figures in different styles, and this might cause male expert participants to find easier to criticize or to empathize with the male dancer whereas female expert participants focus more on female dancers. This indicates that within a ranking or interpretation task, participants yield their focus on a video based on their knowledge.

Sometimes couples can be afraid of falling behind of the rhythm and they start dancing in a rush. Experts can detect if a couple is dancing in a rush by comparing ideal form of the basic figure with the couples' basic figure. If the couple jumps into another move before completing the basic figure, experts say that the couple is in a hurry. However, being in a hurry misled some of the novice participants as if the couples dancing in a rush are dancing much faster than other couples.

Couples' relationship with the audience is a very strong factor for novice participants whereas expert participants focus more on the technique. Costume, hair, and makeup are more likely to be mentioned by novice participants.

Table 4.10 shows the top 15 words of novice and expert groups with their corresponding Tf-Idf score for the Turkish dataset.

Table 4. 10 Feature extraction for Turkish Dataset

Novice Keywords	Translation	Expert Keywords	Translation
sahne 0.348	floor	basmak 0.355	stepping on
anlamak 0.267	understanding	footwork 0.331	footwork
gövde 0.241	body	duygu 0.213	emotion
eğlenmek 0.241	having fun	diz 0.189	knee
yürümek 0.214	walking	dağınık 0.189	messy
son 0.214	end	roll 0.166	roll
değilmi 0.214	Is it	müzikalite 0.166	musicality
bulmak 0.214	finding	bounce 0.166	bounce
selamlamak 0.187	Taking a bow	ayrı 0.166	separate
şey 0.16	thing	sürekli 0.142	continuously

uzatmak 0.134	extending	stiff 0.142	stiff
sevmek 0.134	loving	presentation 0.142	presentation
hâkim 0.134	savant	kırık 0.142	bent
estetik 0.134	estetic	inanılmaz 0.142	unbelievable
şimdi 0.107	now	düşük 0.142	low

Five of the top 15 words of the expert group are English. “Presentation” and “footwork” are evaluation criteria, “bounce” is one of the base moves on which the figures are built up in Samba, “roll” is a basic figure. The word “stiff” was used by expert participants as an adjective referring to couples that cannot provide required upper body moves.

English dataset was created from people’s respond to a set of dance sport competition videos of a dance sport performer and transcript of a dance sport lecture of a former champion. Table 4.11 shows the top 15 words with corresponding Tf-Idf scores from the English dataset.

Table 4. 11 Feature extraction for English Dataset

Intermediate Level of Experience (reddit.com) Keywords	Former World Champion Lecture (youtube.com) Keywords
Rib 0.224	question 0.368
hard 0.204	lecture 0.266
emotion 0.204	quarter 0.245
focus 0.183	gonna 0.245
cha 0.183	cross 0.245
arm 0.183	normally 0.184
level 0.163	final 0.184
definitely 0.163	uh 0.143
fix 0.143	togetherness 0.143
connect 0.143	priority 0.143
chasse 0.143	world 0.123
novice 0.122	today 0.123
lead 0.122	quick 0.123
lack 0.122	lesson 0.123
improve 0.122	idea 0.123

“quarter”, “cross”, and “togetherness” can be considered as more technical terms within the top 15 keywords of the former world champion’s lectures whereas “rib”, “chasse”, and “lead” are the technical terms that have been used by Reddit.com commentators.

As these two are quite different concepts related to same domain, our model might have tracked of the words that are indicators of the concept rather than experience.

CHAPTER 5

CONCLUSION

5.1. Conclusion

The motivation of this study is underpinning of experts' decision making and developing a model that employs the reasoning behind experts' judging process. We hypothesized that it is possible to develop a computational modeling approach by analyzing linguistic utterances from expert and novice evaluators such that the model achieves higher than the chance factor. To test this hypothesis, we have used Multinomial Naïve Bayes (MNB) Classifier (Tf-Idf) and DistilBERT classifier

In this study, Multinomial Naïve Bayes and DistilBERT classifiers were built up and compared to detect experts and novices in dance sport domain with an evaluation task. English and Turkish datasets were fed to the same MNB models. Turkish dataset was fed to dbmdz Distilled Turkish BERT model while English dataset was fed to dbmdz/bert-base-cased-finetuned-conll03-english model.

Turkish data were collected from a group of 22 people. They have been categorized with clustering based on their answers to Likert questions about dance experience. English dataset was created by combining comments to dance sport videos from a group of people who are subscribers of a dance sport page on reddit.com (intermediate level of dancing experience), and transcript of dance sport lecturers from youtube.com (highest level of dancing experience). Regarding the categories, both of the datasets are quite balanced and thus sampling methods were not needed. However, since our datasets are too small, there was a need for increasing the test size. Datasets were shaped in two different ways to get over this problem. First, the cells were divided into cells that have equally distributed number of words. This dataset is called as "balanced" dataset. Secondly, the dataset was shaped in a way that every cell under "Text" column has one sentence each. The second dataset is called as "sentences" dataset.

The data has been pre-processed before it was fed to transporters and vectorizers. Pre-processing consists of lowering letters, removing stopwords & punctuation, and lemmatizing. Zeyrek lemmatizer and wordnet lemmatizer were used for Turkish and English datasets respectively. Zeyrek lemmatizer's output was a list of tuples. Hence, to achieve the desired format for the output of Zeyrek a line of code was employed. The final output was checked and fixed manually.

TfidfVectorizer from Sci-Kit Learn was employed, and different feature selection methods were used to find out the best model for the datasets. Max_features, max_df, and SelectPercentile parameters were configured separately for MNB while corresponding precision, recall and accuracy scores values were noted. For DistilBERT model, learning rate was the parameter that has been manipulated among the trials. The trials were made with both "sentences" dataset and "balanced" dataset. All in whole among 18 trials of MNB and 6 trials of DistilBERT, the highest accuracy score was of 97% for the Turkish "balanced" dataset where learning rate was set to 5e-5. It should be noted that the best score of Turkish balanced

dataset at MNB was 94% when max_features was set to 1000, which is very closed to that of DistilBERT. The highest score for the English dataset was achieved as 95% in MNB classifier with “balanced” dataset where max_features was set to 500 and 1000.

As expected, novice group expressed their thoughts about the couples without using technical terms about dance sport. They focused more on the dresses, gestures, and make ups. The rankings made by novice group were diverging whereas, expert group’s rankings were more consistent and compatible. People in the expert group focused more on the technique and rhythm of the couples during the data collection rather then the outlooks of the couples.

In this thesis study, we investigated if it is possible to distinguish experts and novices from text data on a dance sport evaluation task. We concluded that by studying linguistic utterances from expert and novice evaluators, it is possible to construct a computational modeling strategy that outperforms the chance factor. Among all 24 trials, the optimum model that works accurately on both of the datasets is MNB classifier where max_features is set to 1000. DistilBERT model that is Turkish cased with dbmdz Distilled Turkish BERT was the best option within all trials. However, it should be noted that different results can be achieved as test group and train group are rearranged by manipulating the random state.

5.2. Limitations of the Study & Future Work

In the phase of collecting and establishing the datasets, we had to make some assumptions. While converting the Turkish speech data to text data, we determined the ending of sentences assuming the participant decided to end the sentence when they stop talking for a while. While categorizing the English dataset, we assumed that the subscribers of DanceSport channel on reddit.com cannot be novice because of the technical terms they use in the comments. Thus, we decided to categorize them as “intermediate level of experience”. The video transcripts of former champions were categorized as ”the highest level of experience”.

Natural language processing studies are generally made on much larger datasets. If the dataset volume is increased, more accurate and sensitive models can be achieved. Batch size and epoch parameters in DistilBERT were not changed throughout the trials. Since the computer used in this study was not qualified enough to run higher batch sizes, we stick to the common threshold which is 16. We have manually set the parameters. To find out the best model, parameter optimization can be employed for the future work.

REFERENCES

Alhamzeh, A., Bouhaouel, M., Egyed-Zsigmond, E., & Mitrović, J. (2021). DistilBERT-based argumentation retrieval for answering comparative questions. *CEUR Workshop Proceedings*, *2936*, 2319–2330.

Bar-Hillel, Y. (1964). A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation. *The Present Status of Automatic Translation of Languages*, *1*(1960), 158–163.

Borko, H., & Livingston, C. (1989). Cognition and Improvisation: Differences in Mathematics Instruction by Expert and Novice Teachers. *American Educational Research Journal*, *26*(4), 473–498. https://doi.org/10.3102/00028312026004473

Brill, E., & Mooney, R. J. (1997). Overview of empirical natural language processing. *AI Magazine*, *18*(4), 13–24.

Carey, K., Moran, A., & Rooney, B. (2019). Learning choreography: An investigation of motor imagery, attentional effort, and expertise in modern dance. *Frontiers in Psychology*, *10*(MAR), 1–11. https://doi.org/10.3389/fpsyg.2019.00422

Clarridge, P. B., & Berliner, D. C. (1991). Perceptions of student behavior as a function of expertise. *Journal of Classroom Interaction*, *26*(1), 1–8.

Clermont, C. P., Borko, H., & Krajcik, J. S. (1994). Comparative study of the pedagogical content knowledge of experienced and novice chemical demonstrators. *Journal of Research in Science Teaching*, *31*(4), 419–441. https://doi.org/10.1002/tea.3660310409

Debarnot, U., Sperduti, M., Di Rienzo, F., & Guillot, A. (2014). Experts bodies, experts minds: How physical and mental training shape the brain. *Frontiers in Human Neuroscience*, *8*(MAY), 1–17. https://doi.org/10.3389/fnhum.2014.00280

Endsley, M. R. (1988). Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 97–101. https://doi.org/10.1177/154193128803200221

Engle, R. W., & Bukstel, L. (1978). Memory Processes among Bridge Players of Differing Expertise. *The American Journal of Psychology*, *91*(4), 673. https://doi.org/10.2307/1421515

Gürbüz, T. (2018). 2018 Wei International Academic Conference Proceedings. *Analysis of Interaction of Criteria Affecting a Competitive Dance Couple's Performance in Dance Sport*, *3179*, 266–273.

Henley, M. K. (2014). Is perception of a dance phrase affected by physical movement training and experience? *Research in Dance Education*, *15*(1), 71–82. https://doi.org/10.1080/14647893.2013.835124

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, *28*(1), 127–138. https://doi.org/10.1016/S0364-0213(03)00065-X

Hogan, T., Rabinowitz, M., & Craven, J. A. (2003). Representation in Teaching: Inferences from Research of Expert and Novice Teachers. *Educational Psychologist*, *38*(4), 235–247. https://doi.org/10.1207/S15326985EP3804_3

Housner, L. D., & Griffey, D. C. (1985). Teacher Cognition: Differences in Planning and Interactive Decision Making Between Experienced and Inexperienced Teachers. *Research Quarterly for Exercise and Sport*, *56*(1), 45–53.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). *Memory Influences Subjective Experience : Noise Judgments*. *14*(2), 240–247.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious Influences of Memory for a Prior Event. *Personality and Social Psychology Bulletin*, *13*(3), 314–336. https://doi.org/10.1177/0146167287133003

K. Phillips, J., Klein, G., & R. Sieck, W. (2004). Expertise in Judgment and Decision Making: A Case for Training Intuitive Decision Skills. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (Vol. 1, pp. 297–315). Blackwell Publishing Ltd.

Leinhardt, G., & Greeno, J. G. (1986). The Cognitive Skill of Teaching. *Journal of Educational Psychology*, *78*(2), 75–95. https://doi.org/10.1037/0022-0663.78.2.75

Mack, M. (2020). Exploring cognitive and perceptual judgment processes in gymnastics using essential kinematics information. *Advances in Cognitive Psychology*, *16*(1), 34–44. https://doi.org/10.5709/acp-0282-7

Mann, D. T. Y., Williams, A. M., Ward, P., & Janelle, C. M. (2007). *Perceptual-Cognitive Expertise in Sport : A Meta-Analysis*. 457–478.

Mercier, H., & Heiniger, S. (2018). *Judging the Judges: Evaluating the Performance of International Gymnastics Judges*. 1–14. http://arxiv.org/abs/1807.10021

Merom, D., Cumming, R., Mathieu, E., Anstey, K. J., Rissel, C., Simpson, J. M., Morton, R. L., Cerin, E., Sherrington, C., & Lord, S. R. (2013). *Can social dancing prevent falls in older adults ? a protocol of the Dance , Aging , Cognition , Economics (DAnCE) fall prevention randomised controlled trial*.

Mohd Sanad Zaki Rizvi. (2019). What is BERT | BERT For Text Classification. *Analyticsvidhya*. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Oflazer, K. (2014). Turkish and its challenges for language processing. *Language Resources and Evaluation*, *48*(4), 639–653. https://doi.org/10.1007/s10579-014-9267-2

Ofli, F., Erzin, E., Yemez, Y., & Tekalp, A. M. (2012). Learn2Dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, *14*(3 PART 2), 747–759. https://doi.org/10.1109/TMM.2011.2181492

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Palliser-Sans, R., & Rial-Farràs, A. (2021). *HLE-UPC at SemEval-2021 Task 5: Multi-Depth DistilBERT for Toxic Spans Detection*. 960–966. https://doi.org/10.18653/v1/2021.semeval-1.131

Pedregosa, F., Varoquaux, G., Gramfort, A., & Michel, V. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, *12*, 2825–2830. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

Peskin, J. (1998). Constructing meaning when reading poetry: An expert-novice study. *Cognition and Instruction*, *16*(3), 235–263. https://doi.org/10.1207/s1532690xci1603_1

Pio Di Tore, Alfredo(University of Salerno, I., & Raiola, Gaetano(MIUR Campania, I. (2018). Situation Awarenes in Sports Science: Beyond the Cognitive Paradigm. *INTERNATIONAL SCIENTIFIC JOURNAL OF KINESIOLOGY*, *11*(1), 44–48.

Poulton, E. C. (1957). ON PREDICTION IN SKILLED MOVEMENTS. *PSYCHOLOGICAL BULLETIN*, *54*(6), 467–478.

Raiola, G. (2017). Motor learning and teaching method. *Journal of Physical Education and Sport*, *17*(5), 2239–2243. https://doi.org/10.7752/jpes.2017.s5236

Russo, G., & Ottoboni, G. (2019). The perceptual – Cognitive skills of combat sports athletes: A systematic review. *Psychology of Sport and Exercise*, *44*(April), 60–78. https://doi.org/10.1016/j.psychsport.2019.05.004

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2–6. http://arxiv.org/abs/1910.01108

Seidel, T., Schnitzler, K., Kosel, C., Stürmer, K., & Holzberger, D. (2021). Student Characteristics in the Eyes of Teachers: Differences Between Novice and Expert Teachers in Judgment Accuracy, Observed Behavioral Cues, and Gaze. *Educational Psychology Review*, *33*(1), 69–89. https://doi.org/10.1007/s10648-020-09532-2

Sietas, M. (WDSF). (2015). *Judging System 2.1*. https://www.worlddancesport.org/Rule/Competition/General/Judging_Systems

Silvia, T., & Alexandru Adrian, N. (2016). Study regarding the need To objectify evaluation in Latin-American dances. *Ovidius University Annals, Physical Education and Sport/Science, Movement and Health Series*, *16*(2), 706–710. http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=118881581&S=R&D=s3h&EbscoContent=dGJyMNLe80SeqLE4v%2BvlOLCmr1CeprJSsa%2B4SLCWxWXS&ContentCustomer=dGJyMPGvsEyuqrVJuePfgeyx43zx

Standley, J. M., & Madsen, C. K. (1991). An Observation Procedure to Differentiate Teaching Experience and Expertise in Music Education. *Journal of Research in Music Education*,

39(1), 5–11. https://doi.org/10.2307/3344604

Ste-Marie, D. M. (1999). Expert-Novice Differences in Gymnastic Judging: An Information-processing Perspective. *Applied Cognitive Psychology*, *13*(3), 269–281. https://doi.org/10.1002/(sici)1099-0720(199906)13:3<269::aid-acp567>3.0.co;2-y

Ste-Marie, D. M., & Lee, T. D. (1991). Prior Processing Effects on Gymnastic Judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 126–136. https://doi.org/10.1037/0278-7393.17.1.126

Sun, B., Xiao, W., Feng, X., Shao, Y., Zhang, W., & Li, W. (2020). Behavioral and brain synchronization differences between expert and novice teachers when collaborating with students. *Brain and Cognition*, *139*(December 2019), 105513. https://doi.org/10.1016/j.bandc.2019.105513

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. https://doi.org/10.1016/j.inffus.2016.10.004

WDSF. (2013). *The Judging System 2.0*. https://www.worlddancesport.org/News/WDSF/The_Judging_System_2.0-1130

WDSF. (2015). *Judging Systems*. https://www.worlddancesport.org/Rule/Competition/General/Judging_Systems

World DanceSport Federation, Judging System. (n.d.). https://www.worlddancesport.org/Rule/Competition/General/Judging_Systems

APPENDICES

APPENDIX A Codes of Classifiers

English Preprocessing

```
from nltk.corpus import stopwords
import sys
import re
import nltk.data
from nltk import pos_tag_sents
from nltk.sentiment.util import mark_negation
from nltk import pos_tag
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet as wn
from nltk import pos_tag
from nltk.tokenize import word_tokenize

# relevant packages

from nltk.tokenize import RegexpTokenizer

tokenizer = RegexpTokenizer(r'\w+')

df = pd.read_excel("Eng - Reddit - Youtube_Balanced.xlsx")
df['Text'] = df['Text'].str.lower()
import nltk
nltk.download('averaged_perceptron_tagger')

def tokenize_postag(text):
    tokenized = tokenizer.tokenize(text)
    postagged = nltk.pos_tag(tokenized)
    return postagged

#tokenize and postagging functions

stop_words = stopwords.words('english')
sw_list = ["one", "two", "three", "four", "five", "six", "seven", "eight", "hi", "hello",
"applause", "[", "]", "the", '...' , '--' , """ , '``' , "'s", "a", "of","get", "<lb>", "i", "ve", "t",
"s","m","http","com","lb"]
#we also deleted the counting numbers for samba as we don't want to use counting as an
indicator of experience

stop_words.extend(sw_list)
len(stop_words)
```

```
df['Text-Stop'] = df['Text'].apply(lambda x: ' '.join([word for word in x.split() if word not in
(stop_words)]))
df["postagged"]= df["Text-Stop"].apply(tokenize_postag)

from nltk import pos_tag
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet as wn

lemmatizer = WordNetLemmatizer()

def get_lemmatizer_pos(pos):
    pos_start = pos[0] # Takes the first letter to simplify the POS tag
    if pos_start == "J":
        return wn.ADJ
    elif pos_start == "V":
        return wn.VERB
    elif pos_start == "R":
        return wn.ADV
    else:
        return wn.NOUN


def lemmatize_text(text):

        return [lemmatizer.lemmatize(token[0], pos=get_lemmatizer_pos(token[1])) for token
in text]


# lemmatize function

df['lemmatized'] = df["postagged"].apply(lemmatize_text)
df['lemmatized'] = df['lemmatized'].apply(lambda x: [item for item in x if item not in
stop_words])

#delete the lemmatized words involved in stopwords

def listToString(list):
    str1 = " "
    return (str1.join(list))

#join tokens

df['lemmatized_str'] = df["lemmatized"].apply(listToString)

# apply lemmatization

df[["ID","TEXT","LEMMATIZED_STR","CATEGORY"]].TO_EXCEL("LEMMA_ENGDATA.XLSX")
```

Same steps were repeated for “Eng - Reddit - Youtube_Sentences.xlsx” and had the output “lemma_ENGdata-sentences excel.xlsx”

Preprocessing Turkish

Importing Libraries

```
import matplotlib.pyplot as plt
import nltk
import numpy as np
import pandas as pd
import seaborn as sns
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib notebook
from sklearn.linear_model import LogisticRegression
import sklearn.model_selection
import sklearn.preprocessing as preproc
from sklearn.feature_extraction import text
import pickle
import warnings
warnings.filterwarnings("ignore")

df = pd.read_excel("Turkish-Sentences.xlsx")


def clean_text(text, remove_stopwords = True):
    “removing stopwords'"

    # Convert words to lower case
    text = text.lower()



    # Format words and remove unwanted characters
    text = re.sub(r'https?:\/\/.*[\r\n]*', '', text, flags=re.MULTILINE)
    text = re.sub(r'\<a href', ' ', text)
    text = re.sub(r'&amp;', '', text)
    text = re.sub(r'[_"\-;%()|+&=*%.,!?:#$@\[\]/]', ' ', text)
    text = re.sub(r'<br />', ' ', text)
    text = re.sub(r'\'', ' ', text)

    # remove stop words
    if remove_stopwords:
        text = text.split()
        stops = set(stopwords.words("turkish"))
        text = [w for w in text if not w in stops]
        text = " ".join(text)

    # Tokenize each word
    text =  nltk.WordPunctTokenizer().tokenize(text)
```

```
    return text

import nltk
nltk.download('stopwords')

df['Text_Cleaned'] = list(map(clean_text, df.Text))
```

Lemmatizer

```
!pip install zeyrek
import zeyrek
import nltk
nltk.download('punkt')
analyzer = zeyrek.MorphAnalyzer()
df["analyzed"] = df["Text"].apply(lambda x:analyzer.lemmatize(x))
df["selected"] = df["analyzed"].apply(lambda x: x[:][1][:])
def func1(frame):
    for text in frame:
        return text

df["selected"] = df["analyzed"].apply(lambda x: [x[1][0] for x in x])

df.to_ excel("Turkish-Lemmatized-Sentences.xlsx")
```

Same steps repeated for "Turkish – Balanced.xlsx" and had an output "ID-LemmText-Category (Genişletilmiş Versiyon).xlsx"

Codes of MNB Classifier

Required packages

```
import matplotlib.pyplot as plt

import nltk

import numpy as np

import pandas as pd

import seaborn as sns

import re

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

from sklearn.feature_extraction.text import CountVectorizer

import matplotlib.pyplot as plt

import seaborn as sns
```

```
%matplotlib notebook

from sklearn.linear_model import LogisticRegression

import sklearn.model_selection

import sklearn.preprocessing as preproc

from sklearn.feature_extraction import text

import pickle
import warnings
warnings.filterwarnings("ignore")
import sys
import nltk.data
from nltk.sentiment.util import mark_negation
```

Reading our dataset

```
df = pd.read_excel("ID-LemmText-Category (Genişletilmiş Versiyon).xlsx")

X = df.iloc[:,-2]
y = df.iloc[:,-1]
```

Tf-Idf Vectorizer (max_feature)

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(sublinear_tf=True, max_features = 1000 , lowercase=False , n
gram_range=(1,2))
```

Tf-Idf Vectorizer (max_df)

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(sublinear_tf=True, max_df = 0.3 , lowercase=False , ngram_ra
nge=(1,2))
```

Tf-Idf Vectorizer (SelectPercentile)

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(sublinear_tf=True, lowercase=False , ngram_range=(1,2))

selector = SelectPercentile(f_classif, percentile=20)
selector.fit(vec_train_data, train_label)
vec_train_data = selector.transform(vec_train_data).toarray()
vec_test_data = selector.transform(vec_test_data).toarray()
```

Train Test Split

```
from sklearn.model_selection import train_test_split
train_data , test_data , train_label , test_label = train_test_split(X , y , test_size = 0.2 ,rando
m_state = 0)
```

train data has been vectorized with tfidf

```
vec_train_data = vectorizer.fit_transform(train_data)
vec_train_data = vec_train_data.toarray()
```

test data has been vectorized with tfidf

vec_test_data = vectorizer.transform(test_data).toarray()

matrixes have been converted to dataframe
training_data = pd.DataFrame(vec_train_data , columns=vectorizer.get_feature_names())
testing_data = pd.DataFrame(vec_test_data , columns= vectorizer.get_feature_names())

Importing MNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score,classification_report
clf = MultinomialNB()
clf.fit(training_data, train_label)
y_pred = clf.predict(testing_data)
print(classification_report(test_label , y_pred))

Same steps repeated for "lemma_ENGdata.xlsx"**,** "Turkish-Lemmatized-Sentences.xlsx", "lemma_ENGdata-sentences excel.xlsx".

Codes of DistilBERT Classifier

DISTILBERT ENGLISH
!PIP INSTALL TRANSFORMERS

IMPORT PANDAS AS PD
IMPORT TENSORFLOW AS TF
IMPORT TRANSFORMERS
FROM TRANSFORMERS IMPORT DISTILBERTTOKENIZER
FROM TRANSFORMERS IMPORT TFDISTILBERTFORSEQUENCECLASSIFICATION
FROM SKLEARN.MODEL_SELECTION IMPORT TRAIN_TEST_SPLIT

MODEL_NAME = "DBMDZ/BERT-BASE-CASED-FINETUNED-CONLL03-ENGLISH" # THIS MODEL IS A CHECKPOINT SAVED FOR ENGLISH CASED DATASETS

BATCH_SIZE = 16
N_EPOCHS = 3 # INCREASING THE EPOCH WILL LEAD TO OVERFITTING, IT IS YOUR CHOICE

Splitting train, test, and validation data
import pandas as pd
data = pd.read_excel('lemma_ENGdata-sentences excel.xlsx' *) #English*

data.dropna(inplace=True)

X_train_base, X_val_base = train_test_split(data, test_size=25)
X_train_base, X_test_base = train_test_split(X_train_base, test_size=120)
X_train_base.index = range(0,len(X_train_base))
X_val_base.index = range(0,len(X_val_base))
X_test_base.index = range(0,len(X_test_base))

X_train =X_train_base['Lemmatized Text']
X_test =X_test_base['Lemmatized Text']
y_train = X_train_base['Category']
y_test = X_test_base['Category']

Maximum length

MAX_LEN = X_train.apply(lambda s: len([x for x in s.split()])).max()
MAX_LEN = MAX_LEN if MAX_LEN < 512 else 512

DistilBERT Tokenizer
tokenizer = DistilBertTokenizer.from_pretrained(MODEL_NAME)

train_encodings = tokenizer(list(X_train.values), truncation=True, padding=True)
test_encodings = tokenizer(list(X_test.values), truncation=True, padding=True)

print(f'First paragraph: \'{X_train[:1]}\')
print(f'Input ids: {train_encodings["input_ids"][0]}')
print(f'Attention mask: {train_encodings["attention_mask"][0]}')

Turn labels and encodings into a tf.Dataset object

train_dataset = tf.data.Dataset.from_tensor_slices((dict(train_encodings),
list(y_train.values)))

test_dataset = tf.data.Dataset.from_tensor_slices((dict(test_encodings),
list(y_test.values)))

Fine-tuning with native TensorFlow
model = TFDistilBertForSequenceClassification.from_pretrained(MODEL_NAME)

optimizerr = tf.keras.optimizers.Adam(learning_rate=5e-5)
losss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True) # Computes the cr ossentropy loss between the labels and predictions.
model.compile(optimizer=optimizerr,
loss=losss,
metrics=['accuracy'])

history = model.fit(train_dataset.shuffle(len(X_train)).batch(BATCH_SIZE),
epochs=N_EPOCHS,
batch_size=BATCH_SIZE)

Evaluation of the Model
model.evaluate(test_dataset.shuffle(len(X_test)).batch(BATCH_SIZE), return_dict=True, bat ch_size=BATCH_SIZE)

DISTILBERT TURKISH
!PIP INSTALL TRANSFORMERS

IMPORT PANDAS AS PD
IMPORT TENSORFLOW AS TF
IMPORT TRANSFORMERS
FROM TRANSFORMERS IMPORT DISTILBERTTOKENIZER
FROM TRANSFORMERS IMPORT TFDISTILBERTFORSEQUENCECLASSIFICATION
FROM SKLEARN.MODEL_SELECTION IMPORT TRAIN_TEST_SPLIT

MODEL_NAME = 'DBMDZ/DISTILBERT-BASE-TURKISH CASED'# THIS MODEL IS A CHECKPOINT SAVED FOR TURKISH CASED DATASETS

Splitting train, test, and validation data

```
import pandas as pd
data = pd.read_excel("Turkish-Lemmatized-Sentences.xlsx") #Turkish

data.dropna(inplace=True)

X_train_base, X_val_base = train_test_split(data, test_size=25)
X_train_base, X_test_base = train_test_split(X_train_base, test_size=120)
X_train_base.index = range(0,len(X_train_base))
X_val_base.index = range(0,len(X_val_base))
X_test_base.index = range(0,len(X_test_base))

X_train =X_train_base['Lemmatized Text']
X_test =X_test_base['Lemmatized Text']
y_train = X_train_base['Category']
y_test = X_test_base['Category']
```

Maximum length

```
MAX_LEN = X_train.apply(lambda s: len([x for x in s.split()])).max()
MAX_LEN = MAX_LEN if MAX_LEN < 512 else 512

DistilBERT Tokenizer
tokenizer = DistilBertTokenizer.from_pretrained(MODEL_NAME)

train_encodings = tokenizer(list(X_train.values), truncation=True, padding=True)
test_encodings = tokenizer(list(X_test.values), truncation=True, padding=True)

print(f'First paragraph: \'{X_train[:1]}\'')
print(f'Input ids: {train_encodings["input_ids"][0]}')
print(f'Attention mask: {train_encodings["attention_mask"][0]}')
```

Turn labels and encodings into a tf.Dataset object

```
train_dataset = tf.data.Dataset.from_tensor_slices((dict(train_encodings),
                                        list(y_train.values)))

test_dataset = tf.data.Dataset.from_tensor_slices((dict(test_encodings),
                                        list(y_test.values)))

Fine-tuning with native TensorFlow
model = TFDistilBertForSequenceClassification.from_pretrained(MODEL_NAME)


optimizerr = tf.keras.optimizers.Adam(learning_rate=5e-5)
```

```
losss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True) # Computes the
crossentropy loss between the labels and predictions.
model.compile(optimizer=optimizerr,
              loss=losss,
              metrics=['accuracy'])

history = model.fit(train_dataset.shuffle(len(X_train)).batch(BATCH_SIZE),
          epochs=N_EPOCHS,
          batch_size=BATCH_SIZE)
Evaluation of the Model
model.evaluate(test_dataset.shuffle(len(X_test)).batch(BATCH_SIZE), return_dict=True,
batch_size=BATCH_SIZE)
```

Same steps repeated for "ID-LemmText-Category (Genişletilmiş Versiyon).xlsx"

APPENDIX B Feature Extraction

```
df = pd.read_excel("feature extraction ENG.xlsx")
X = df.iloc[:,-2] #Text sütunu X
y = df.iloc[:,-1] # Category sütunu y

from sklearn.feature_extraction.text import CountVectorizer
import re
docs=df['Text'].tolist()

import nltk
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('english')

cv=CountVectorizer(max_df=0.85,stop_words=stopwords)
word_count_vector=cv.fit_transform(docs)

from sklearn.feature_extraction.text import TfidfTransformer
tfidf_transformer=TfidfTransformer(smooth_idf=True,use_idf=True)
tfidf_transformer.fit(word_count_vector)

feature_names=cv.get_feature_names()
doc=docs[0]
tf_idf_vector=tfidf_transformer.transform(cv.transform([doc]))




def sort_coo(coo_matrix):

    tuples = zip(coo_matrix.col, coo_matrix.data)

    return sorted(tuples, key=lambda x: (x[1], x[0]), reverse=True)

sorted_items=sort_coo(tf_idf_vector.tocoo())
```

Get the feature names and tf-idf score of top 10 items

```
def extract_topn_from_vector(feature_names, sorted_items, topn=10):
```

Extract only the top 10

```
keywords=extract_topn_from_vector(feature_names,sorted_items,10)
```

```
def sort_coo(coo_matrix):
    tuples = zip(coo_matrix.col, coo_matrix.data)
    return sorted(tuples, key=lambda x: (x[1], x[0]), reverse=True)

def extract_topn_from_vector(feature_names, sorted_items, topn=10):
    """get the feature names and tf-idf score of top n items"""

    #use only topn items from vector
    sorted_items = sorted_items[:topn]

    score_vals = []
    feature_vals = []

    # word index and corresponding tf-idf score
    for idx, score in sorted_items:

        #keep track of feature name and its corresponding score
        score_vals.append(round(score, 3))
        feature_vals.append(feature_names[idx])

    #create a tuples of feature,score
    #results = zip(feature_vals,score_vals)
    results= {}
    for idx in range(len(feature_vals)):
        results[feature_vals[idx]]=score_vals[idx]

    return results
```

```
the document that we want to extract keywords from
doc=docs[1]
#1 for novices 0 for experts
#generate tf-idf for the given document
tf_idf_vector=tfidf_transformer.transform(cv.transform([doc]))

#sort the tf-idf vectors by descending order of scores
sorted_items=sort_coo(tf_idf_vector.tocoo())

#extract only the top 15
keywords=extract_topn_from_vector(feature_names,sorted_items,15)
# now print the results
print("\n=====Doc=====")
print(doc)
print("\n===Keywords===")
for k in keywords:
    print(k,keywords[k])
```

APPENDIX C Results From the Datasets

Results From The English Dataset

Table C. 1 English Dataset MNB, SP=10

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.96	0.97	0.80	0.69	0.87	0.81	30	52
Category 1	0.83	0.74	0.97	0.98	0.89	0.84	30	46
accuracy					0.88	0.83	60	89
Macro avg.	0.89	0.86	0.88	0.84	0.88	0.83	60	98
Weighted avg.	0.89	0.86	0.88	0.83	0.88	0.82	60	98

Table C. 2 English Dataset MNB, SP=20

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.95	1.00	0.67	0.69	0.78	0.82	30	52
Category 1	0.74	0.74	0.97	1.00	0.84	0.85	30	46
accuracy					0.82	0.84	60	98
Macro avg.	0.85	0.87	0.82	0.85	0.81	0.84	60	98
Weighted avg.	0.85	0.84	0.82	0.84	0.81	0.83	60	98

Table C. 3 English Dataset MNB, max_df=0.3

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.95	0.98	0.67	0.65	0.78	0.79	30	52
Category 1	0.74	0.72	0.97	0.98	0.84	0.84	30	46
accuracy					0.82	0.82	60	98
Macro avg.	0.85	0.85	0.82	0.82	0.81	0.81	60	98

Weighted avg.	0.85	0.87	0.82	0.82	0.81	0.81	60	98

Table C. 4 English Dataset MNB, max_df=0.5

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.96	1.00	0.73	0.63	0.83	0.78	30	52
Category 1	0.78	0.71	0.97	0.97	0.87	0.83	30	46
accuracy					0.85	0.81	60	98
Macro avg.	0.87	0.85	0.85	0.80	0.85	0.80	60	98
Weighted avg.	0.87	0.86	0.85	0.81	0.85	0.80	60	98

Table C. 5 English Dataset MNB, max_features=300

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.88	0.93	0.83	0.93	0.85	30	52
Category 1	0.93	0.82	0.93	0.87	0.93	0.84	30	46
accuracy					0.93	0.85	60	98
Macro avg.	0.93	0.85	0.93	0.85	0.93	0.85	60	98
Weighted avg.	0.93	0.86	0.93	0.85	0.93	0.85	60	98

Table C. 6 English Dataset MNB, max_features=500

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.94	0.94	0.97	0.85	0.95	0.89	30	52
Category 1	0.97	0.84	0.93	0.93	0.95	0.89	30	46
accuracy					0.95	0.89	60	98
Macro avg.	0.95	0.89	0.95	0.89	0.95	0.89	60	98
Weighted avg.	0.95	0.89	0.95	0.89	0.95	0.89	60	98

Results from the Turkish Datasets

Table C. 7 Turkish Dataset MNB, SP=5

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	1.00	0.69	0.76	0.63	0.87	0.66	17	73
Category 1	0.78	0.68	1.00	0.73	0.88	0.70	14	78
accuracy					0.87	0.68	31	151
Macro avg.	0.89	0.68	0.88	0.68	0.87	0.68	31	151
Weighted avg.	0.90	0.68	0.87	0.68	0.87	0.68	31	151

Table C. 8 Turkish Dataset MNB, SP=20

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	1.00	0.77	0.71	0.68	0.83	0.72	17	73
Category 1	0.73	0.73	1.00	0.81	0.85	0.77	14	78
accuracy					0.84	0.75	31	151
Macro avg.	0.87	0.75	0.85	0.75	0.84	0.75	31	151
Weighted avg.	0.88	0.75	0.84	0.75	0.84	0.75	31	151

Table C. 9 Turkish Dataset MNB, max_df=0.3

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.77	0.93	0.77	0.93	0.77	14	73
Category 1	0.94	0.78	0.94	0.78	0.94	0.78	17	78
accuracy					0.94	0.77	31	151
Macro avg.	0.93	0.77	0.93	0.77	0.93	0.77	31	151
Weighted avg.	0.94	0.77	0.94	0.77	0.94	0.77	31	151

Table C. 10 Turkish Dataset MNB, max_df=0.5

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.87	0.77	0.93	0.77	0.90	0.77	14	73
Category 1	0.94	0.78	0.88	0.78	0.91	0.78	17	78
accuracy					0.90	0.77	31	151
Macro avg.	0.90	0.77	0.91	0.77	0.90	0.77	31	151
Weighted avg.	0.91	0.77	0.90	0.77	0.90	0.77	31	151

Table C. 11 Turkish Dataset MNB, max_features=300

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.76	0.69	0.87	0.60	0.81	0.64	16	73
Category 1	0.86	0.67	0.75	0.74	0.80	0.70	15	78
accuracy					0.81	0.68	31	151
Macro avg.	0.81	0.68	0.81	0.67	0.81	0.67	31	151
Weighted avg.	0.81	0.68	0.81	0.68	0.81	0.67	31	151

Table C. 12 Turkish Dataset MNB, max_features=500

	Precision Balanced	Precision Sentences	Recall Balanced	Recall Sentences	Accuracy-Score Balanced	Accuracy-Score Sentences	Support Balanced	Support Sentences
Category 0	0.93	0.75	0.81	0.68	0.87	0.71	16	73
Category 1	0.82	0.73	0.93	0.78	0.87	0.75	15	78
accuracy					0.87	0.74	31	151
Macro avg.	0.88	0.74	0.87	0.73	0.87	0.73	31	151
Weighted avg.	0.88	0.74	0.87	0.74	0.87	0.73	31	151

DistilBERT Results

Table C. 13 DistilBERT, learning rate=5e-4

	Turkish Dataset Balanced	Turkish Dataset Sentences	English Dataset Sentences
Accuracy	0.77	0.51	0.55
Cross-Entropy Loss	0.57	0.69	0.70

APPENDIX D Top Features

Table D. 1 Top 100 Features of the Expert Group With Corresponding Tf-IDf Scores

Feature	Translation	Feature	Translation
basmak 0.355	to step on	açmak 0.118	to open
footwork 0.331	footwork	anlam 0.118	meaning
duygu 0.213	emotion	yeni 0.095	new
diz 0.189	knee	sayı 0.095	number
dağınık 0.189	messy	prominade 0.095	promenade
roll 0.166	roll	problem 0.095	problem
müzikalite 0.166	musicality	connection 0.095	connection
bounce 0.166	bounce	bağlantı 0.095	connection
ayrı 0.166	separate	ancak 0.095	however
sürekli 0.142	continuous	akışkan 0.095	fluid
stiff 0.142	stiff	çeyrek 0.071	quarter
presentation 0.142	presentation	yarım 0.071	half
kırık 0.142	bent	vurgu 0.071	emphasis
inanılmaz 0.142	unbelievable	temiz 0.071	clean
düşük 0.142	low	seviye 0.071	level
partnering 0.118	partnering	sağlam 0.071	durable
lâzım 0.118	required	poz 0.071	pose
klâsik 0.118	classic	kapmak 0.071	capture
dağılmak 0.118	to disperse	düz 0.071	straight
bozulmak 0.118	to be broken down	aksan 0.071	accent
boz 0.047	gray	üç 0.047	three
box 0.047	box	çalmak 0.047	to play
basit 0.047	simple	zolasyon 0.047	isolation
basic 0.047	basic	zigzag 0.047	zigzag
bariz 0.047	obvious	yapı 0.047	structure
ball 0.047	ball	yanyana 0.047	side by side
action 0.047	action	whisk 0.047	whisk
kaba 0.047	rude	zleyiciyle 0.024	with the audience
ilerlemek 0.047	to proceed	zleyici 0.024	the audience
görev 0.047	mission	zaten 0.024	already
frame 0.047	frame	timing 0.047	timing
eğlence 0.047	fun	temel 0.047	basic
düzeltmek 0.047	to fix	tamamen 0.047	completely
dizlemek 0.047	knee action	sürtmek 0.047	to drag
dik 0.047	upright	spor 0.047	sport
boş 0.047	empty	skill 0.047	skill
bozuk 0.047	broke down	savrulmak 0.047	to be dispersed
bozmak 0.047	to break down	run 0.047	run
rotasyon 0.047	rotation	profesyonel 0.047	professional

ritmik 0.047	rhythmical	pozisyon 0.047	position
oyun 0.047	play	abartı 0.047	exaggerate
movement 0.047	movement	özel 0.024	special
metre 0.047	meter	öbür 0.024	the other
merkez 0.047	center	çizmek 0.024	to draw
kurgulamak 0.047	to fictionalize	çirkin 0.024	ugly
korkmak 0.047	to scare	çevik 0.024	agile
kaybetmek 0.047	to lose	çerçeve 0.024	frame
kapamak 0.047	to close	çekiştirmek 0.024	to tug
kalite 0.047	quality	zorlamak 0.024	to force

Table D. 2 Top 100 Features of the Novice Group With Corresponding Tf-IDf Scores

Feature	**Translation**	**Feature**	**Translation**
sahne 0.348	floor	yaşamak 0.053	to live
anlamak 0.267	to understand	yaratmak 0.053	to create
gövde 0.241	body	tatlı 0.053	sweet
eğlenmek 0.241	to have fun	sıradan 0.053	regular
yürümek 0.214	to walk	siyah 0.053	black
son 0.214	the last	seçim 0.053	choice
mi 0.214	is it	parlamak 0.053	to shine
bulmak 0.214	to find	ortalama 0.053	average
selamlamak 0.187	to take a bow	oran 0.053	ratio
şey 0.16	thing	makyaj 0.053	make up
uzatmak 0.134	to prolog	kurmak 0.053	to establish
sevmek 0.134	to love	keyif 0.053	joy
hâkim 0.134	having good command (of a subject)	karşılık 0.053	correspondence
estetik 0.134	esthetic	istek 0.053	will
şimdi 0.107	now	irite 0.053	irritating
pantolon 0.107	pants	herhâlde 0.053	maybe
ne 0.107	what	gülümsemek 0.053	to smile
kafa 0.107	head	gülmek 0.053	to smile
heralde 0.107	maybe	görüyorum 0.053	I see
gibi 0.107	like	göremedim 0.053	I couldn’t see
devam 0.107	continuo	giyinmek 0.053	to wear
birinci 0.107	the first	final 0.053	final
yakınmak 0.08	complain	falan 0.053	so-and-so
tâbi 0.08	of course	etek 0.053	skirt
saç 0.08	hair	dair 0.053	regarding
sadece 0.08	only	bel 0.053	waist
performans 0.08	performance	altı 0.053	under
ileri 0.08	forward	şaşırmak 0.027	to be surprised
hata 0.08	error	şalap 0.027	messy (movement)
doğru 0.08	correct	ısınmak 0.027	to warm up
dolamak 0.08	to wind	çinden 0.027	from inside
arka 0.08	back	çeşitli 0.027	various
öteki 0.053	the other	çekingen 0.027	shy
önem 0.053	importance	çekimser 0.027	abstainer
çıkarmak 0.053	to eject	çekilmek 0.027	to withdraw
çekişmek 0.053	to conflict	çağrıştırmak 0.027	to connotate
çarpmak 0.053	to crash	âdeta 0.027	almost
yormak 0.053	to make tired	zlerken 0.027	while watching

yoksa 0.053	or	ziyade 0.027	rather than
yiler 0.053	good (they are)	zerafet 0.027	grace
zannetmek 0.027	to suppose	varlık 0.027	existence
yuvarlak 0.027	circular	uzaklaşmak 0.027	to move away
yumuşak 0.027	soft	tür 0.027	specie
yorum 0.027	comment	tutulmak 0.027	to catch on
yapay 0.027	artificial	tutuk 0.027	hesitant
yanında 0.027	next to	tutku 0.027	passion
yalap 0.027	messy (movement)	tercih 0.027	choice
yakışmak 0.027	to be suitable	ten 0.027	skin
vurgulamak 0.027	to emphasise	temas 0.027	contact
vasat 0.027	average/fair	teklemek 0.027	to stutter