Essays on Retail Operations and The Recent Pandemic (COVID-19): Using Mathematical and Text-Mining Approaches

By MARYAM MAHDIKHANI

A dissertation submitted to the

Graduate School—Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Dr. Yao Zhao

and approved by

_____

_____

_____

_____

_____

Newark, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

Essays on Retail Operations and The Recent Pandemic (COVID-19): Using Mathematical and Text-Mining Approaches

by Maryam Mahdikhani

Dissertation Director : Dr. Yao Zhao

This dissertation consists of three essays. The first essay examines on auction design and the last two essays apply sentiment analysis methodologies on big data. The first paper of my dissertation examines the auction design with negative externality and its impact on the optimal mechanism design. In light of previous studies, our research shows that auctioning a good may impact the seller's payoff and those who lose the object. We simplify the potential mechanism by depriving buyers of their right to absolute non-participation. Our characterizations are thus tailored towards understanding bidders' type space, and the information structure of single-object auctions with negative externality's set up.

The second paper of my dissertation aims to predict helpful reviews on Amazon Fashion products and identify the most frequent terms in such reviews. We

choose features from topics using the latent Dirichlet allocation (LDA) model and topics plus Bi-grams using the TF-IDF vectorizer. We then use the features to enhance the performance of support vector machine (SVM) classifier to predict the helpfulness of reviews. The research is performed on a large corpus of Amazon fashion reviews. We find that reviews gets more votes when they are more specific regarding quality of product and return experience.

The third essay of my dissertation is motivated by tweets on COVID-19 and the retweeting behavior. Our research objective is to predict tweet's popularity based on the volume of retweets regardless of the user's followers. We examine the features selection, including (i) topics by using LDA, (ii) N-grams by using TF-IDF vectorizer, and (iii) topics plus Bi-grams TF-IDF vectorizer. We use the extracted features on Random Forest (RF) classifier, SVM classifier, and Logistic Regression (LR) classifier. We find that RF has the highest accuracy for predicting the volume of retweets by particularly using topics plus Bi-grams TF-IDF vectorizer.

Acknowledgments

First and foremost I would like to express my special appreciation and thanks to my advisor Dr. Yao Zhao. It has been an honor to be his Ph.D. student. He has taught me to think about research ideas and stories behind them. I am also grateful to him for his valuable support and guidance during my Ph.D. journey. Without his trust and consistent support, I could not overcome the many crisis situations during my Ph.D. studies.

I would like to thank to Dr. Soo Hyun Cho, and Dr. Shaoqiong Zhao, who continuously found time to discuss the research with me and guide me. They helped me on my third and fourth chapters in this dissertation. I want to express my appreciation to Dr. Jian Yang for his help on my second chapter. I am deeply grateful to him for his support and valuable feedback.

I would like to thank Dr. Mark Rodgers for serving on the committee and supporting me in improving my dissertation with his insightful comments. In addition, I am also grateful to all my friends especially to Aziza Jones, Arim Park, and Mahak Nagpal who have helped me throughout the years at Rutgers. Most importantly, none of my achievement would have been possible without the love and patient of my family. Particularly, for my husband, Christopher for holding my hand during the crises and for my parents for being patient not seeing me since

I started my Ph.D. Words cannot describe how grateful I am for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online retail sales have steadily increased from 13.2% in 2017 to 14.14% in 2018 and 16% in 2019[1]. An extensive variety of available products and, fast and free shipping options are the main reasons for the growth of online shopping. This growth, however, leads to some difficulties. Despite all the efforts of online retailers (ORs) to facilitate online shopping, consumers are still confronted with a few problems while making online purchases. The biggest predicament is that consumers are unable to try products firsthand and learn about them before making a purchase. Since the product's quality is not determined until consumers receive it, returning online purchases is quite a common phenomenon.

To maintain a convenient shopping and return experience, many ORs have started to invest in opening physical stores to facilitate online shopping and decrease the hassle for customers. Additionally, some ORs have begun acquiring other retailers in their industry via auctions to expand their market share in different channels, which influences the seller's and other bidders' payoff, as explained in my first essay.

[1]"E-commerce sales surpassed 10% of total retail sales in 2019 for the first time." *Business Insider* Feb 24, 2020.

In my first essay, we study the auction design with negative externality, wherein the buyers' externalities influence the seller's payoff. Our study is motivated by Amazon's acquisition of Whole Foods and its effects on other groceries' operations[2]. When Amazon made an offer to buy Whole Foods, other potential competitors, including Walmart, Kroger, and delivery companies, started to make bids for it as well [3]. There were three different types of bidders in the auction: aggressive buyers, like Amazon that was willing to bid at higher price to win the object; strategic buyers, like Walmart that was willing to bid at certain price by considering certain constrains regarding its strategy; and potential growth buyers, like delivery companies that were willing to bid to extend their business with respect to their budget. The bidder relieve their information and their type space to the seller based on their strategy of acquiring the object. Each bidder has certain value for the object and they do not relieve their true value to the seller but they are aware of the negative externality that other bidder might cause for them.

The winner can significantly increase its benefit by revealing its identity and harming the loser. Therefore, the seller should identify its actual value in the decision mechanism, and the optimal information structure should be intended on promote the best bid, which may not necessarily be the highest price. We simplified the potential mechanism by depriving buyers of their right to absolute non-participation. The objective of our study is to help the seller identify an optimal mechanism. Our characterizations are tailored towards understanding the bidders' type space, and the information structure of a single-object auction

[2]"Amazon to Buy Whole Foods for $13.4 Billion"*New York Times* June 16, 2017.
[3]"Walmart? Amazon may find rival bidders for Whole Foods" *US Today* July 22, 2017.

with negative externalities' set up. We show that if the negative externalities created by the sale are higher than the seller's payoff, then the seller is better off not auctioning the object away.

In the second essay of my dissertation, we discuss the importance of reviews' helpfulness for future purchases from the perspectives of both consumers and retailers. Our research is motivated by Amazon's voting system for reviews, wherein the helpful reviews are defined as those with more than three helpful votes. We aim to predict helpful reviews and investigate the most frequent terms in such reviews. We select features from topics by using latent Dirichlet allocation (LDA) model and topics plus bi-grams by using term frequency–inverse document frequency (TF-IDF) analysis to obtain consumers' feelings, expressed through helpful and non-helpful reviews. We then use the features to enhance the performance of the support vector machine (SVM) classifier to predict the reviews' helpfulness. We demonstrate our model's performance in prediction accuracy by comparing the two categories of features on SVM model. The models are applied to a large corpus of Amazon fashion review text bodies and they predict the helpfulness of reviews using sentiment analysis techniques. For both approaches, we assess the performance impact of creating a training set that includes not only the rating system (i.e., one to 5-star reviews), but also votes on the helpfulness of reviews. Moreover, using an available data set of Amazon fashion reviews, for each iteration, we perform classification experiments on samples from different product categories.

This method tends to be more accurate than other methods since we train the classifier using real-world data sets. Furthermore, we use a test set to determine

the accuracy of the system and applied cross-validation to validate the results. Moreover, we seek to identify the most frequent terms in helpful reviews with respect to polarity of reviews. Furthermore, we aim to understand the possible reasons of getting more votes on the reviews. We examine whether the length of reviews has an impact on its helpfulness. We find that longer reviews are not considered to be as helpful from consumers' perspective on fashion field. Consumers are more likely to trust and vote for reviews that explain the quality of the product and, in case of a misfit product, the return experience.

In the third and the last essay of my dissertation, we examine the tweets related to COVID-19 pandemic and the importance of the spread of content of the tweet during the pandemic. In this study, we aim to predict the popularity of tweets based on the volume of retweets. We categorize the dataset into popular tweets with higher than 136 times retweets or equal to 136 times retweets and non-popular tweets with less than 136 times retweets. We use the topics analysis by LDA for the short text and add co-occurrence terms of network by using TF-IDF vectorizer to extract the features and obtain users' feeling and information related to pandemic. Furthermore, we compare the different category of features such as (i) topics analysis (by using LDA), (ii) n-grams analysis (by using TF-IDF vectorizer) including; uni-gram TF-IDF vectorizer, bi-grams TF-IDF vectorizer, and tri-grams TF-IDF vectorizer, and (iii) topics plus bi-grams TF-IDF vectorizer. We applied the aforementioned categories of features on three supervised machine learning algorithms including Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). We find that RF has the highest accuracy compared

to other classifier and also among all the features, topics analysis plus bi-grams TF-IDF vectorizer improves the accuracy of classifier significantly. We check the validation of our models by using cross-validation with five-folds and compare the results. The performance of models are also checked by using Receiver Operating Characteristics (ROC) with cross-validation. Moreover, in terms of exploratory analysis, we find that United States is the most active country on the Twitter during the pandemic, and the popular hashtag is still coronavirus beside all the other events that occurred during pandemic.

Chapter 2

Auction involving Externalities

2.1 Introduction

Nowadays, store chains and Brick and Mortar Retailers (B&MRs) are facing several challenges to maintain their market share and compete with online retailers (ORs). B&MR are trying to make strategic alliances by joining big and well-known competitors in the online channel or physical stores. Improving the efficiency of operation, and capturing new consumers are some of the main reasons for retailers to cite acquisitions. The goal of increasing sales among stronger competitors with strategic behavior raises several challenges for the retailers to acquire their competitor. Therefore, the problem of allocating limited resources among strategic users with private information is often addressed through the framework of auctions or mechanism design. In our case, the more influential retailer as a seller is the limited strategic resource among the weaker retailers as buyers who want to maximize their firms' values and gain a competitive edge in their industry field by getting an object. Therefore, buyers compete to acquire the object by considering their value on the object and the seller's value on the object. For instance, Target

acquired Shipt to improve the efficiency of their delivery service [1], and Walmart bought jet.com to capture consumers who are willing to purchase directly online[2].

There is a large body of literature on mechanism design, and specifically on auction design that is restricted to modeling uncertainty by a single parameter for the public good, or several private goods by considering the quantity. In line with previous research by Jehiel et al. (1999), there is a possibility that auctioned goods might have an influence on the auction's participants who did not win the auction, and the outcome of the auction affects their future business.

There are several examples of this situation in the real world, such as changes of ownership in competitive markets, the patent sale, and several cases like that. In order to avoid negative externalities, many bidders overpaid for an auctioned good, which conclude to the winner curse problem. The negative externality hypothesis generally shows that bidders are not only willing to pay in an auction because of their value toward the object, but also to reduce the negative externality (e.g., Jehiel and Moldovanu (1996); Jehiel et al. (1999)).

Our case study highlights the consequences of Amazon's Whole Foods acquisition on payoffs for the winner by itself and other bidders. Amazon.com Inc announced it would acquire grocery store chain Whole Foods Market for $ 13.7 billion in 2017 [3]. In 2017, Amazon made an offer to buy Whole Foods, and other potential competitors decided to make a bid for it as well. Some of the buyers wanted to raise the price for the powerful bidder. The buyers in Whole Foods acquisition were including : Walmart, Kroger, Small chains (many regional players are privately

[1]"Why Target bought delivery startup Shipt" *digitalcommerece360* March 20, 2018.
[2]"Walmart's acquisition of Jet.com ..."businessinsider June 13, 2019.

held and buying Whole Foods could give them a more national profile like Wegmans and Trader Joe's), Foreign chains, Target, Costco, Sprouts, and Delivery companies [4]. Buyers do not necessarily have beliefs about their own or other buyer's values. The winner can increase its benefit significantly in the long run, which harms other buyer's payoff. In the case of Amazon's acquisition of Whole Foods, the winner can influence the market share and change the grocery operations. The purpose of this paper is to examine that the seller's profit can be influenced by the externalities that the winner causes on other bidders.

Furthermore, the bid has also impact on the auction design and it can influence the valuation of rival buyers. In previous research, there is a framework where buyers have private information on the externalities they cause to others, and they assumed these externalities do not depend on the loser's identity. Our paper contributed to the previous researches by considering the seller's problem with the externalities that the winner causes on other's outcomes. This study organized as follows: we give a background about auction theory with externalities in §2.2, and we introduce our model in §2.3. In §2.3 , we have three subsections that start with research question in §2.3.1, follows with §2.3.2, and ends with §2.3.3. We have the analysis of a special case in §2.4, and continue with the question on will various symmetries help? in §2.5. We have a case study analysis in section §2.6, and we have the conclusion at §2.7. The proofs are provided on Appendix.

[4]"Could there be a bidding war for Whole Foods?" *CNN* June 19, 2017.

2.2 Literature Review

A large body of literature related to auction theory is on developing different models to explain the bidder values. The private values model explains that each bidder has a private value for the object and does not impact other bidders' value. The popular value model explains that all the bidders can have the same value for the object, but they do not share information about it before the sale. Moreover, some models have a combination of private and common models as an especial case. In all of these cases, they consider the mechanisms that maximize seller's profit as an optimal mechanism. Our research contributes to two streams of literature: (i) auction design, particularly the growing literature on the effect of payoff function on the outcome of market by having the externality, (ii) single object auction model. In our research, we seek to link the seller's optimal payoff models in economics literature to the literature related to the models of negative externalities. Therefore, in our framework, the buyers have valuations for getting the item, and they also have negative externality valuations when losing the item to their competitors.

Existing literature concerning the auction theory concentrated mostly on the case where the item being auctioned is valuable only to the bidder who possesses it. Many papers proposed models regarding to the information rather than the payoff structure; see, e.g., Riley and Samuelson (1981), Myerson (1981), and Milgrom and Weber (1982). Early research related to auction theory can be found in McAfee and McMillan (1987); Rothkopf and Harstad (1994); Wilson (1992). Lorentziadis (2016) has the main contribution in biding from a game theory perspective by

highlighting the impact of auction theory in practice.

Maskin and Riley (2000) studied on pure strategy equilibrium in the auction where buyers with lower values and higher bids can participate in a Vickrey auction. Krishna (2009) explained the bidder with lower values are forced to bid higher to stay in the competition compared to the bidders with higher values. In our case study, Amazon is considered as a powerful bidder (i.e., the bidder with a higher value), therefore, the powerful bidder bids more aggressively when he competes against a strong opponent rather than a weak one. In this literature, each buyer perceives all other buyers as direct competitors, which brings the specific stream of research named models with externalities.

To the best of our knowledge,Jehiel and Moldovanu (1996) pioneered the consideration of different payoffs to buyers when the identity of the auction winner changes. In the initial study, they studied the complete-information case involving negative externalities and an incomplete-information case in which buyers have private information on the externalities they cause to others. In a study complementary to the latter case, Jehiel et al. (1999) allowed players to have private information on the payoff to itself when the item is in either seller or any other's possession. Here, Jehiel and Moldovanu (1996); Jehiel et al. (1999) seemed to have equated minimizing buyers' surpluses with maximizing the seller's revenue.

Aseff and Chade (2008) focused on the case of two units, positive externalities, and buyers' payoff functions that are recognized in types and externality parameters. Varma (2002) also considered the type of externalities and examined each buyer's equilibrium willingness to pay depends on the identities of her opponents. Ettinger

(2003) studied a case where the loser of the auction is sensitive about the winner's bid.

The second research stream is related to single object auction, which is the vast body of literature in auction theory design literature. Varma (2002) considered an auction for a single individual object, and found the buyer's payoffs depend on the winner's identity in a single-good environment. Furthermore, Varma and Lopomo (2010) examined both dynamic and sealed-bid auctions and their bidding behavior when the winner reveals the information at a single object auction. Papadimitriou and Pierrakos (2011) studied the optimal auction design under incentive-compatible constrain by focusing on single-item auction. The most recent research is conducted by Bei et al. (2019) where the problem of revenue maximization in the single-item auction is examined within the robust framework. They find that by increasing the number of bidders, the optimal auction's format does not have a significant impact in a single-item auction.

Common features of all these studies are that the dominant strategy in auction design is implementable, and maximizes the payoff for the seller by giving away the single object to the winner, whether there is a negative externality or not. We provide a characterization of auction design in which the optimal welfare is influenced by bidding strategy and all bidders' types, and the negative externalities of losers affect the seller's payoff.

2.3 The Model

2.3.1 Research Questions

Most of the existing literature has addressed the question of payoff gains from acquisitions. Many researchers defined several possible scenarios that would occur if a firm is affected by competitor's acquisition and how bidders would behave when they enter to an auction and how they end up having the item while they overpaid for it. This ambiguity happens because of two situations; firstly, the bidder is not aware of other existing bidders, and their willingness to buy the item, so the bidder has to avoid the cost of losing the object.

Therefore, the bidder tries to acquire the item, since losing the item is even more costly and brings negative externalities. Secondly, the bidder knows the value of the acquisition for herself and other potential competitors, which can change the type of industry or bring negative externalities for other bidders.

To fix the idea, consider the following example related to Amazon's acquisition of Whole Foods when Amazon convinced Whole Foods not to involve other bidders in $13.7 billion cash. Furthermore, there is another example with same situation where Versace also was sold to United State label Michael Kors for $2 billion [5]. The difference between these two examples is, in the former example, the industry type changed by involving online channels into the food industry, while in the second example, they expended the size of their market share by adding more brands in their company. In this research, we examine the appropriate answer for the following question: how can the seller auction off the item to a group of

interested buyers to have an optimal welfare? The most critical challenge the auctioneer faces is; not being certain about the value of the buyers for a item, which describes how much they are willing to pay for a item, and how much they are willing to share their information for the sellers and other buyers.

Therefore, the decision mechanism needs to enforce the buyers to share reliable information in order to have optimal welfare. The optimization problem, in this case, is more complicated since it is not clear how rational bidders will play. Previous researches cope with this uncertainty of how buyers play by considering mechanisms where rational bidders are willing to tell the seller their complete type. Such mechanisms are considered to be incentive compatible (IC), where the bidder share their true type regardless of other bidders' type. We have this constraint in our model, and we explain it in detail in the following sections.

Furthermore, even after restricting the research space to IC auctions, it is still a very difficult problem to solve if no prior is known over the bidder's types. There are many solutions in the literature by adopting a Bayesian viewpoint, considering that a prior does exist and is known for both the seller and the buyers, and targeting the optimal expected welfare. Myerson (1981) studied how to capture an optimal revenue by considering the case where bidders are single-dimensional. Although after Myerson's work, a large proportion of literature studied about multi-dimensional problem (i.e., the setting where the bidders may have different values for the item), we are still far from an optimal mechanism.

Our focus on this work is to fill this important gap in the mechanism design literature by analyzing a special case and studying various symmetries on our

model to examine how it helps. Given that our study is filling the gap in optimal mechanism design, we examine how traditional auction differs from our design and whether it still works.

2.3.2 Basic Setup

Designate the seller as player 0 and the buyers as players 1 to n. For player $i = 0, 1, ..., n$, the reward to her will be v^i_j when it is player $j = 0, 1, .., n$ that wins the item. Note player 0 winning the item just means that the item has not been auctioned away. If buyer i pays the seller x and the item is won by player j, she will obtain payoff

$$v^i_j - x. \tag{2.1}$$

If the buyers' payments to the seller forms a vector $\mathbf{x} \equiv (x^i)_{i=1,...,n}$ and the item is eventually won by player j, the seller will obtain payoff

$$v^0_j + \sum_{i=1}^{n} x^i. \tag{2.2}$$

There is a subset $\mathcal{V}$ of $\Re^{(n+1)\times(n+1)}$ that contains all possible payoff profiles. The space $\mathcal{V}$ describes the auction's payoff structure. If every $v^i_j = 0$ for every $\mathbf{V} \equiv (v^i_j)_{i,j=0,1,...,n} \in \mathcal{V}$, we would have reverted back to the traditional case without externalities. As for information structure, we can have many varieties to choose from, just like in the case of the traditional auction. When every player $i = 1, ..., n$ submits her bid x_i, the vector $x \equiv (x_i)_{i=1,...,n}$ is the basis on which player 0 will make decisions. Regardless, we can let $\mathcal{T}_i$ be each player i's private type space,

and let $\mathcal{T}_{n+1}$ be the space that take care of residual valuation uncertainties left uncovered by all players' private information. For convenience, let $\mathcal{T} \equiv \prod_{i=0}^{n} \mathcal{T}_i$ be the space of all type profiles. When $\mathbf{t} \equiv (\mathbf{t}_i)_{i=0,1,...,n} \in \mathcal{T}$ and $\mathbf{t}_{n+1} \in \mathcal{T}_{n+1}$ are given, we suppose players' valuations of the item is given by some $\tilde{\mathbf{V}}(\mathbf{t}, \mathbf{t}_{n+1}) \equiv (\tilde{v}^i_j(\mathbf{t}, \mathbf{t}_{n+1}))_{i,j=0,1,...,n}$, where $\tilde{\mathbf{V}}$ is a mapping from $\mathcal{T} \times \mathcal{T}_{n+1}$ to $\Re^{(n+1)\times(n+1)} \cup \{\infty\}$ where ∞ is an $(n+1) \times (n+1)$-dimensional vector whose every component is the one-dimensional ∞.

If no player knows anything, $\mathcal{T}$ would contain just one default point say $\bar{\mathbf{t}} \equiv (\bar{\mathbf{t}}_i)_{i=0,1,...,n}$. When players' combined knowledge can always pinpoint their entire payoff profile, $\mathcal{T}_{n+1}$ would contain one default point $\bar{\mathbf{t}}_{n+1}$. In the case of complete information, we can further equate every $\mathcal{T}_i$ with $\mathcal{V}$ and let $\tilde{\mathbf{V}}(\mathbf{t}, \bar{\mathbf{t}}_{n+1})$ be any $\mathbf{t}_i$ when $\mathbf{t}_1 = \mathbf{t}_2 = \cdots = \mathbf{t}_n$ while ∞ when some $\mathbf{t}_i \neq \mathbf{t}_j$.

Without loss of generality, we restrict attention to revelation mechanisms in which every buyer i reports to the seller her true type $\mathbf{t}_i$. Then, when it comes the turn for the seller to make her decision, she would have learned the entire type profile $\mathbf{t} \equiv (\mathbf{t}_i)_{i=0,1,...,n}$. Let

$$\Delta_n \equiv \left\{ \mathbf{p} \equiv (p_j)_{j=1,...,n} \in [0,1]^n : \sum_{j=1}^{n} p_j \leq 1 \right\}. \tag{2.3}$$

It can be treated as the space of all probabilistic assignments of the item to buyers. The seller's decision can be summarized as some $(\mathbf{p}, \mathbf{x})$, where $\mathbf{p}$ is a mapping from $\mathcal{T}$ to Δ_n and $\mathbf{x} \equiv (x^i)_{i=1,...,n}$ a mapping from $\mathcal{T}$ to $\Re^n$. Given any type profile $\mathbf{t} \in \mathcal{T}$, the probabilistic-assignment vector $\mathbf{p}(\mathbf{t}) \equiv (p_j(\mathbf{t}))_{j=1,...,n}$ would contain the

chances that the item is to be given to buyers j; also, $p_0(\mathbf{t}) \equiv 1 - \sum_{j=1}^{n} p_j(\mathbf{t})$ would be the chance that the seller is to keep the item; in addition, the payment vector $\mathbf{x}(\mathbf{t}) \equiv (x^i(\mathbf{t}))_{i=1,\ldots,n}$ would contain the payments that buyer i must make to the seller.

Compared to Jehiel and Moldovanu (1996) and Jehiel et al. (1999), we have simplified the potential mechanism by depriving buyers of their rights to absolute non-participation. Note the "external options" afforded by a buyer's non-participation, even it is strictly enforced, amount to the item being assigned to other buyers or left with the seller. These are describable by the probabilistic assignments $\mathbf{p}$ already. Also, we believe a buyer has ample opportunity to signal to the seller her unwillingness to participate so actively that it can rarely be non-binding in real life.

For instance, in a traditional first-price or second-price auction where the seller retains the right to keep the item, a buyer who does not feel like to participate could bid \$1 billion below her valuation. If she still ends up with the item plus the nearly \$1 billion compensation, it is a pure reflection of others' equal unwillingness as well as the seller's utter disgust against the item. To this buyer, however, there seems to be hardly any loss. Of course, the revelation principle has relieved us of the task to explicitly model the mappings from buyers' types to their signals.

Suppose the seller adopts some decision mechanism $(\mathbf{p}, \mathbf{x})$. Then according to (2.1), by reporting her type as $\mathbf{s}_i$ a buyer i with private type $\mathbf{t}_i$ will face the following expected payoff when other players together report truthfully their type

profile $\mathbf{t}_{-i} \equiv (\mathbf{t}_j)_{j \neq i} \in \mathcal{T}_{-i} \equiv \prod_{j \neq i} \mathcal{T}_j$:

$$u^i(\mathbf{s}_i, \mathbf{t}_i, \mathbf{t}_{-i}; \mathbf{p}, \mathbf{x}) \equiv \sum_{j=0}^{n} p_j(\mathbf{s}_i, \mathbf{t}_{-i}) \cdot \tilde{v}^i_j(\mathbf{t}_i, \mathbf{t}_{-i}) - x^i(\mathbf{s}_i, \mathbf{t}_{-i}). \tag{2.4}$$

If all buyers report their types truthfully, the seller's expected payoff under each type profile $\mathbf{t} \in \mathcal{T}$ would follow from (2.2) to be

$$u^0(\mathbf{t}; \mathbf{p}, \mathbf{x}) \equiv \sum_{j=0}^{n} p_j(\mathbf{t}) \cdot \tilde{v}^0_j(\mathbf{t}) + \sum_{i=1}^{n} x^i(\mathbf{t}). \tag{2.5}$$

Given any payoff structure $\mathcal{V}$ and information structure, our job is to find a mechanism $(\mathbf{p}, \mathbf{x})$ that optimizes the seller's expected payoff while ensuring all buyers' incentive compatibility.

A simpler case is when for $i = 1, ..., n$, we fix $t_{i0}(x) = x_i p_i(x)$ whereas $t_{ij}(x) = 0$ for $j = 1, ..., i-1$. This is the case where each x_i is the payment to be made by player i to player 0, and no other side payments are involved. With this simplification, the only decision that player 0 has to make is the winner assignment mapping $p : \Re^n \to \Delta_{n+1}$. Because of our more complicated payoff structure, this can already be regarded as a generalization to the traditional first-price auction. For some value vector $v \equiv (v_i)_{i=0,1,...,n}$, our setup can be reduced to the latter by letting $v^i_i = v_i$ and $v^i_j = 0$ for $i \neq j$.

2.3.3 Information Structure

We suppose each private-type space $\mathcal{T}_i$ is a multi-dimensional interval $[\underline{\mathbf{t}}_i, \bar{\mathbf{t}}_i]$ in some $\Re^{\bar{d}_i}$. For convenience, let $\underline{\mathbf{t}} = (\underline{\mathbf{t}}_i)_{i=1,...,n}$ and $\bar{\mathbf{t}} = (\bar{\mathbf{t}}_i)_{i=1,...,n}$. Now $\mathcal{T} = \prod_{i=1}^{n} \mathcal{T}_i$

is just the interval $[\underline{\mathbf{t}}, \overline{\mathbf{t}}]$. We can also define $\underline{\mathbf{t}}_{-i}$ and $\overline{\mathbf{t}}_{-i}$ and equate $[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]$ with $\mathcal{T}_{-i}$. The valuation-matrix space $\mathcal{V}$ can be understood as the range space of the mapping $\tilde{\mathbf{V}}$ from $[\underline{\mathbf{t}}, \overline{\mathbf{t}}]$ to $\Re^{(n+1)\times(n+1)}$:

$$\mathcal{V} \equiv \left\{ \tilde{\mathbf{V}}(\mathbf{t}) \in \Re^{(n+1)\times(n+1)} : \mathbf{t} \in [\underline{\mathbf{t}}, \overline{\mathbf{t}}] \right\}. \tag{2.6}$$

Let there be a strictly positive and continuous probability density function (pdf) f on the space $[\underline{\mathbf{t}}, \overline{\mathbf{t}}]$. For each $i = 1, ..., n$, we can define the marginal pdf f_i as

$$f_i(\mathbf{t}_i) \equiv \int_{[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]} d\mathbf{t}_{-i} \cdot f(\mathbf{t}_i, \mathbf{t}_{-i}), \qquad \forall \mathbf{t}_i \in [\underline{\mathbf{t}}_i, \overline{\mathbf{t}}_i]. \tag{2.7}$$

We can also define the conditional pdf $f_{-i}(\cdot, \cdot|\cdot)$ as in

$$f_{-i}(\mathbf{t}_{-i}|\mathbf{t}_i) \equiv \frac{f(\mathbf{t}_i, \mathbf{t}_{-i})}{f_i(\mathbf{t}_i)}, \tag{2.8}$$

for every $\mathbf{t}_i \in [\underline{\mathbf{t}}_i, \overline{\mathbf{t}}_i]$ and $\mathbf{t}_{-i} \in [\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]$. While in its incomplete-information form, the nuclear-weapons sales model of Jehiel and Moldovanu (1996), could be understood as the following particular case. Here, the seller possesses no prior private information and so we can let $\mathbf{t} \equiv (\mathbf{t})_{i=1,...,n}$. In other words, the seller is indifferent of who owns the item, nor does the buyers derive pleasure or pain from her owning the item; therefore, $\tilde{v}^0_0(\mathbf{t}) = \tilde{v}^0_1(\mathbf{t}) = \cdots = \tilde{v}^0_n(\mathbf{t}) = \tilde{v}^1_0(\mathbf{t}) = \cdots = \tilde{v}^n_0(\mathbf{t}) = 0$. In addition, every buyer i knows and only knows her own valuation and the negative externality that her possession of the item would cause other buyers; hence, each $\mathbf{t}_i = (\pi_i, \alpha_i)$ and each $\tilde{v}^i_j(\mathbf{t})$ is merely a function of

$\mathbf{t}_j \equiv (\pi_j, \alpha_j)$; it is equal to π_j when $j = i$ and to $-\alpha_j$ when $j \neq i$. For $n = 3$, a matrix representation of the $\tilde{v}^i_j(\mathbf{t})$'s, with rows i, columns j, and $\mathbf{t} \equiv (\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3) \equiv ((\pi, \alpha_1), (\pi_2, \alpha_2), (\pi_3, \alpha_3))$, is as follows:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \pi_1 & -\alpha_2 & -\alpha_3 \\ 0 & -\alpha_1 & \pi_2 & -\alpha_3 \\ 0 & -\alpha_1 & -\alpha_2 & \pi_3 \end{pmatrix}. \tag{2.9}$$

If the above could be considered as the case where buyers' information is "column-wise" here, then in the model of Jehiel et al. (1999), buyers' information became "row-wise". Here, payoffs to the seller when the item is auctioned off to some buyer are reflected in a commonly known vector $(\bar{t}^0_j)_{j=1,\ldots,n}$. For the type profile $\mathbf{t} \equiv (\mathbf{t}_i)_{i=1,\ldots,n} \equiv ((t^i{}_j)_{j=1,\ldots,n})_{i=1,\ldots,n}$ where each $t^i{}_j$ is buyer i's gain when buyer j gets the item. Again for $n = 3$, a matrix representation of the $\tilde{v}^i_j(\mathbf{t})$'s, with $\mathbf{t} \equiv (\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3) \equiv ((t^1_1, t^1_2, t^1_3), (t^2_1, t^2_2, t^2_3), (t^3_1, t^3_2, t^3_3))$, is as follows:

$$\begin{pmatrix} 0 & t^0_1 & t^0_2 & t^0_3 \\ 0 & t^1_1 & t^1_2 & t^1_3 \\ 0 & t^2_1 & t^2_2 & t^2_3 \\ 0 & t^3_1 & t^3_2 & t^3_3 \end{pmatrix}. \tag{2.10}$$

Suppose players operate under our general information structure while the seller adopts a mechanism $(\mathbf{p}, \mathbf{x})$. Then, we can follow 2.4 to derive buyer i's

expected gain $U^i(\mathbf{s}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x})$ when she reports her type as $\mathbf{s}_i$ while her actual type is $\mathbf{t}_i$; it is equal to

$$\begin{aligned}&\int_{[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]} d\mathbf{t}_{-i} \cdot f^i_{-i}(\mathbf{t}_{-i}|\mathbf{t}_i) \cdot u^i(\mathbf{s}_i, \mathbf{t}_i, \mathbf{t}_{-i}; \mathbf{p}, \mathbf{x}) \\ &\quad = \sum_{j=0}^n \int_{[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]} d\mathbf{t}_{-i} \cdot f^i_{-i}(\mathbf{t}_{-i}|\mathbf{t}_i) \cdot p_j(\mathbf{s}_i, \mathbf{t}_{-i}) \cdot \tilde{v}^i_j(\mathbf{t}_i, \mathbf{t}_{-i}) \\ &\quad\quad - \int_{[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]} d\mathbf{t}_{-i} \cdot f^i_{-i}(\mathbf{t}_{-i}|\mathbf{t}_i) \cdot x^i(\mathbf{s}_i, \mathbf{t}_{-i}),\end{aligned} \tag{2.11}$$

where $f^i_{-i}(\cdot|\cdot)$ is defined through 2.7 and 2.8. Following 2.5, we can obtain the seller's expected gain as in

$$\begin{aligned}U^0(\mathbf{p}, \mathbf{x}) &= \int_{[\underline{\mathbf{t}}, \overline{\mathbf{t}}]} d\mathbf{t} \cdot f_{0,1,...,n}(\mathbf{t}) \cdot u^0(\mathbf{t}; \mathbf{p}, \mathbf{x}) \\ &= \sum_{j=0}^n \int_{[\underline{\mathbf{t}}, \overline{\mathbf{t}}]} d\mathbf{t} \cdot p_j(\mathbf{t}) \cdot f_{0,1,...,n}(\mathbf{t}) \cdot \tilde{v}^0_j(\mathbf{t}) + \sum_{i=1}^n \int_{[\underline{\mathbf{t}}, \overline{\mathbf{t}}]} d\mathbf{t} \cdot f_{0,1,...,n}(\mathbf{t}) \cdot x^i(\mathbf{t}).\end{aligned} \tag{2.12}$$

Using random-vector notation where each random $\mathbf{t}_i$ is denoted as $\mathbf{\Theta}_i$, we may simplify 2.11 and 2.12, respectively, into

$$U^i(\mathbf{s}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x}) = \sum_{j=0}^n \mathbb{E}\left[p_j(\mathbf{s}_i, \mathbf{\Theta}_{-i}) \cdot \tilde{v}^i_j(\mathbf{t}_i, \mathbf{\Theta}_{-i})|\mathbf{t}_i\right] - \mathbb{E}\left[x^i(\mathbf{s}_i, \mathbf{\Theta}_{-i})|\mathbf{t}_i\right], \tag{2.13}$$

for $i = 1, ..., n$, and

$$U^0(\mathbf{p}, \mathbf{x}) = \sum_{j=0}^n \mathbb{E}\left[p_j(\mathbf{\Theta}) \cdot \tilde{v}^0_j(\mathbf{\Theta})\right] + \sum_{i=1}^n \mathbb{E}\left[x^i(\mathbf{\Theta})\right]. \tag{2.14}$$

Our objective is to help the seller identify a mechanism $(\mathbf{p}, \mathbf{x})$ that would maximize $U^0(\mathbf{p}, \mathbf{x})$ while maintaining that for every $i = 1, ..., n$ and every $\mathbf{t}_i \in [\underline{\mathbf{t}}_i, \overline{\mathbf{t}}_i]$,

From 2.13 and 2.15, we see that $S^i(\mathbf{t}_i; \mathbf{p}, \mathbf{x})$ is equal to

$$\sup \left\{ \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{s}_i, \boldsymbol{\Theta}_{-i}) \cdot \tilde{v}^i{}_j(\mathbf{t}_i, \boldsymbol{\Theta}_{-i}) | \mathbf{t}_i\right] - \mathbb{E}\left[x^i(\mathbf{s}_i, \boldsymbol{\Theta}_{-i}) | \mathbf{t}_i\right] : \mathbf{s}_i \in [\underline{\mathbf{t}}_i, \overline{\mathbf{t}}_i] \right\}. \tag{2.16}$$

2.4 Analysis of a Special Case

We consider the special case is where all the random vectors $\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_n$ are independent of each other. From 2.13, we can see that U^i's dependence on x^i would only be through y^i where $y^i(\mathbf{s}_i) \equiv \mathbb{E}[x^i(\mathbf{s}_i, \boldsymbol{\Theta}_{-i})]$. That is, buyer i's utility would depend on the payment rule applied to herself only through the average around her own reporting. Just because $\tilde{v}^i{}_j$ depends on $\mathbf{t}_{-i}$ as well as $\mathbf{t}_i$, the notion of $q_j(\mathbf{s}_i) \equiv \mathbb{E}[p_j(\mathbf{s}_i, \boldsymbol{\Theta}_{-i})]$ would not be much of a use here.

Furthermore, suppose every $\tilde{v}^i{}_j$ for $i = 1, ..., n$ satisfies

$$\tilde{v}^i{}_j(\mathbf{t}_i, \mathbf{t}_{-i}) = \langle \tilde{\mathbf{a}}^i{}_j(\mathbf{t}_{-i}), \mathbf{t}_i \rangle + \tilde{b}^i{}_j(\mathbf{t}_{-i}), \tag{2.17}$$

where $\tilde{\mathbf{a}}^i{}_j$ is continuous mapping from $[\underline{\mathbf{t}}_{-i}, \overline{\mathbf{t}}_{-i}]$ to $\Re^{\bar{d}_i}$ and $\tilde{b}^i{}_j$ a continuous mapping from the same domain to $\Re$. Here, every buyer i's valuation of the item is affine in her own type $\mathbf{t}_i$. We still allow the coefficients involved in the affine form to depend on the identity j of the player who obtains the item and the other-buyer type $\mathbf{t}_{-i}$. A further special case of this occurs when each $\tilde{\mathbf{a}}^i{}_j(\mathbf{t}_{-i}) = \bar{\mathbf{a}}^i{}_{ji}$ and each $\tilde{b}^i{}_j(\mathbf{t}_{-i}) = \sum_{k \neq i} \langle \bar{\mathbf{a}}^i{}_{jk}, \mathbf{t}_k \rangle + \bar{b}^i{}_j$ for fixed appropriate-dimensional vectors $\bar{\mathbf{a}}^i{}_{jk}$ and

fixed scalars $\bar{b}^i{}_j$. If so, 2.17 would become, for $i \neq 0$,

$$\tilde{v}^i{}_j(\mathbf{t}) = \sum_{k=1}^{n} \langle \bar{\mathbf{a}}^i{}_{jk}, \mathbf{t}_k \rangle + \bar{b}^i{}_j. \tag{2.18}$$

This reflects that all players' types contribute to a player's valuation in linear fashions. In Jehiel and Moldovanu (1996), only the $\bar{\mathbf{b}}^i{}_{jj}$'s are nonzero; whereas, Jehiel et al. (1999), only the $\bar{\mathbf{b}}^i{}_{ji}$'s are.

Plugging 2.17 into 2.13 while noting the independence, we have

$$U^i(\mathbf{s}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x}) = \langle \mathbf{V}^i(\mathbf{s}_i; \mathbf{p}), \mathbf{t}_i \rangle - W^i(\mathbf{s}_i; \mathbf{p}, x^i), \tag{2.19}$$

where

$$\mathbf{V}^i(\mathbf{s}_i; \mathbf{p}) \equiv \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{s}_i, \boldsymbol{\Theta}_{-i}) \cdot \tilde{\mathbf{a}}^i{}_j(\boldsymbol{\Theta}_{-i}) \right], \tag{2.20}$$

and

$$W^i(\mathbf{s}_i; \mathbf{p}, x^i) \equiv \mathbb{E}\left[x^i(\mathbf{s}_i, \boldsymbol{\Theta}_{-i}) \right] - \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{s}_i, \boldsymbol{\Theta}_{-i}) \cdot \tilde{b}^i{}_j(\boldsymbol{\Theta}_{-i}) \right]. \tag{2.21}$$

An exploitable feature of 2.19 is that the term $\langle \mathbf{V}^i(\mathbf{s}_i; \mathbf{p}), \mathbf{t}_i \rangle$, which captures the entirety of $U^i(\mathbf{s}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x})$'s $\mathbf{t}_i$-dependency, is independent of the payment rule $\mathbf{x}$ while being linear and hence convex in $\mathbf{t}_i$. Let $S^i(\mathbf{t}_i; \mathbf{p}, \mathbf{x})$ be buyer i's surplus at type $\mathbf{t}_i$ when she tells the truth under a mechanism $(\mathbf{p}, \mathbf{x})$. Due to 2.15,

$$S^i(\mathbf{t}_i; \mathbf{p}, \mathbf{x}) \equiv U^i(\mathbf{t}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x}) = \sup\left\{ U^i(\mathbf{s}_i, \mathbf{t}_i; \mathbf{p}, \mathbf{x}) : \mathbf{s}_i \in [\underline{\mathbf{t}}_i, \bar{\mathbf{t}}_i] \right\}. \tag{2.22}$$

According to Proposition 1 of Krishna and Maenner (2001), $S^i(\mathbf{t};\mathbf{p},\mathbf{x})$ will be convex function of $\mathbf{t}$ and its value will be determined by $\mathbf{p}$ alone up to an additive constant; hence, $S^i(\cdot;\mathbf{p},y^i)$ has a y^i-independent sub-differential say $\partial S^i(\mathbf{t}_i;\mathbf{p})$ at any $\mathbf{t}_i$. From 2.19,

$$\frac{\partial U^i(\mathbf{s}_i,\mathbf{t}_i;\mathbf{p},y^i)}{\partial \mathbf{t}_i} = \mathbf{Q}^i(\mathbf{s}_i;\mathbf{p}). \tag{2.23}$$

By the envelope theorem, $\mathbf{Q}^i(\mathbf{s}_i^*(\mathbf{t}_i;\mathbf{p},y^i);\mathbf{p}) \in \partial S^i(\mathbf{t}_i;\mathbf{p})$. By 2.15,

$$S^i(\mathbf{t}_i;\mathbf{p},y^i) = U^i(\mathbf{t}_i,\mathbf{t}_i;\mathbf{p},y^i) \quad \text{and hence} \quad \mathbf{s}_i^*(\mathbf{t}_i;\mathbf{p},y^i) = \mathbf{t}_i, \tag{2.24}$$

Thus, $\mathbf{Q}^i(\mathbf{t}_i;\mathbf{p}) \in \partial S^i(\mathbf{t}_i;\mathbf{p})$. While revisiting 2.27,

$$S^i(\underline{\mathbf{t}}_i;\mathbf{p},y^i) = \langle \mathbf{Q}^i(\underline{\mathbf{t}}_i;\mathbf{p}),\underline{\mathbf{t}}_i\rangle + \sum_{j=0}^{n} \mathbb{E}\left[p_j(\underline{\mathbf{t}}_i,\boldsymbol{\Theta}_{-i})\cdot \tilde{b}^i{}_j(\boldsymbol{\Theta}_{-i})\right] - y^i(\underline{\mathbf{t}}_i). \tag{2.25}$$

When 2.20 is viewed in combination with 2.27 to 2.25, we see that the payment function $\mathbf{y}^i$ is completely determined by $\mathbf{y}^i(\underline{\mathbf{t}}_i)$ and the allocation rule $\mathbf{p}$. Note the seller aims at maximizing 2.14. Tapping into Krishna and Maenner (2001), again, we see have the following theorem.

Theorem 1 *We assume for each player i, the type $\boldsymbol{\Theta}_{-i}$ is convex, and $S^i(\mathbf{t_i};\mathbf{p},\mathbf{y^i})$ is a convex function of t_i. Therefore, the expected equilibrium utility function is convex in any incentive compatibility mechanism $(\mathbf{p},\mathbf{y^i})$, therefore, for any smooth path $\mathbf{r}$ joining $\mathbf{t}_1$ to $\mathbf{t}_2$, where $\mathbf{r}$ is an arbitrary smooth path from $\underline{\mathbf{t}}_i$ to $\mathbf{t}_i$, say the*

straight one.

$$S^i(\mathbf{t}_i;\mathbf{p},y^i) = S^i(\underline{\mathbf{t}}_i;\mathbf{p},y^i) + \int_{\underline{\mathbf{t}}_i}^{\mathbf{t}_i} \langle \mathbf{Q}^i(\mathbf{s}_i;\mathbf{p}), d\mathbf{r}(\mathbf{s}_i)\rangle, \tag{2.26}$$

where $\mathbf{Q}^i(\mathbf{s}_i;\mathbf{p})$ *is a subgradient of* S^i *at* t_i*.*

The equation 2.26 corresponds to the importance of $\mathbf{Q}^i(\mathbf{s}_i;\mathbf{p})$ as a subgradient of a convex function S^i. Therefore, we have the following proposition,

Proposition 1 *We explain the action for our mechanism* (p,y) *where* Q^i *responses to conditional probability assignment functions. Therefore, we have incentive compatible constrain for buyer* i *if and only if the vector field* Q^i *is monotone for all* t_i*.*

$$y^i(\mathbf{t}_i) = \langle \mathbf{Q}^i(\mathbf{t}_i;\mathbf{p}), \mathbf{t}_i\rangle - S^i(\mathbf{t}_i;\mathbf{p},y^i) + \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{t}_i,\boldsymbol{\Theta}_{-i})\cdot \tilde{b}^i{}_j(\boldsymbol{\Theta}_{-i})\right]. \tag{2.27}$$

where through 2.26.

The proposition 1 states that for incentive compatible mechanisms, the payment is determined by the type of buyer $y^i(t_i)$. The validity of 2.26, regardless of the path $\mathbf{r}$ being chosen, just means that $\mathbf{Q}^i(\cdot;\mathbf{p})$ is conservative. Just because $S^i(\cdot;\mathbf{p},y^i)$ is convex, we also have monotonicity on its derivatives; thus, for any $\mathbf{t}_i \in [\underline{\mathbf{t}}_i, \overline{\mathbf{t}}_i]$,

$$\langle \mathbf{t}_1\mathbf{Q}^1(\mathbf{t}_1;\mathbf{p}) - \mathbf{Q}^2(\mathbf{t}_2;\mathbf{p})\rangle \geq 0. \tag{2.28}$$

Furthermore, by part 2.26, the payment and surplus of the lowest type $y_i(\underline{\mathbf{t}}_i)$ and

describes the incentive compatible mechanisms, and the $y^i(\mathbf{t_i})$ is uniquely determined by $y_i(\underline{\mathbf{t}}_i)$. Furthermore, we denote the vector $\big(y_1(\underline{\mathbf{t}}_1), \cdots, y_n(\underline{\mathbf{t}}_n)\big)$ by $y(\underline{\mathbf{t}})$.

By considering negative externality case as $\omega_i < 0$ then the most risky problem for buyer i is when seller gives the object to one of i's opponents. Therefore, we have

$$\hat{y}_i(\underline{\mathbf{t}}_i) = \langle \mathbf{Q}^i(\underline{\mathbf{t}}_i; \mathbf{p}), \underline{\mathbf{t}}_i \rangle - \omega_i \tag{2.29}$$

Therefore, we have the following proposition;

Proposition 2 *In the mechanism design where every* $\mathbf{Q}^i(\underline{\mathbf{t}}_i; \mathbf{p})$ *,* $i \in I$*, is monotone. Therefore,* $\hat{y}_i(\underline{\mathbf{t}}_i)$ *is the optimal auction mechanism with probability assignment function* p*.*

Proposition 2 shows that in the case of negative externality, it is enough to check the participation constraint for the critical type. Furthermore, we have a special case for the seller's utility which is as follows:

$$U^0(\mathbf{p}, \mathbf{x}) = \sum_{j=0}^{n} \bar{a}^0_j \cdot \mathbb{E}[p_j(\mathbf{\Theta})] + \sum_{j=0}^{n} \sum_{k=0}^{n} \mathbb{E}[p_j(\mathbf{\Theta}) \cdot \langle \bar{\mathbf{b}}^0_{jk}, \mathbf{\Theta}_k \rangle] + \sum_{i=1}^{n} \mathbb{E}[x^i(\mathbf{\Theta})]. \tag{2.30}$$

2.5 Will Various Symmetries Help?

Suppose all the $\mathcal{T}_i$'s are the same $[\underline{\mathbf{t}}, \overline{\mathbf{t}}] \subset \Re^{\bar{d}}$ for some dimension $\bar{d}$. When each $\mathbf{t}_{-i}$ is arranged in the fashion of $(\mathbf{t}_{i+1}, ..., \mathbf{t}_n, \mathbf{t}_1, ..., \mathbf{t}_{i-1})$, all the $\tilde{\mathbf{a}}^i_i$ vector functions are the same $\tilde{a}^1_1$ and all the $\tilde{b}^i_i$ functions are the same $\tilde{b}^1_1$. When each $\mathbf{t}_{-i}$ is arranged in the fashion of all the $\tilde{a}^i_j$ functions for $i \neq j$ are the same $\tilde{a}^1_2$, , all the $\tilde{b}^i_j$

further symmetry, let Π be the set of all $n!$ permutations of $\{1, 2, ..., n\}$. Also consider one-dimensional bidding strategies. After symmetry, suppose all buyers use a common bidding strategy $b(\cdot)$. This auction maximizes the seller's expected payoff among all standard auctions with one dimensional bids.

Therefore, we define the function Π for each permutation as follows:

$$\Pi(b) := (b_1, ..., b_n)$$

for each $b \in R^N$. In this case, the seller cannot make the outcome depend on the identity of the buyers, because all buyers use the same bidding strategy. Therefore, by considering a standard bidding mechanism (p, y_i), buyer i's conditional payment $y^i(b_i)$ and probability assignment vector $\langle \mathbf{Q}^i(\mathbf{b}_i; \mathbf{p}), \mathbf{t}_i \rangle$ for any bid $b_i \in [\underline{\mathbf{b}}, \overline{\mathbf{b}}]$ are defined by

$$Y^i(\mathbf{b}_i; \mathbf{p}, y_i) \equiv y_i - \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{b}_i, \mathbf{\Theta}_{-i}) \cdot \tilde{b}^i{}_j(\mathbf{\Theta}_{-i})\right].$$

and,

$$\mathbf{Q}^i(\mathbf{b}_i; \mathbf{p}) \equiv \sum_{j=0}^{n} \mathbb{E}\left[p_j(\mathbf{b}_i, \mathbf{\Theta}_{-i}) \cdot \tilde{\mathbf{a}}^i{}_j(\mathbf{\Theta}_{-i})\right],$$

By considering (p, y_i) as an anonymous mechanism, and b_i as a symmetric bidding strategy, we have $y_i \equiv y_1$ for all $i > 1$. Given that, the expected payment function is the same for all buyers. The following lemma also describes that the conditional probability assignment function is the same for all the buyers and all the bids b_i for all the buyers i have the same conditional probability of getting the object. Suppose (P, y_i) is anonymous and b is a symmetric bidding strategy. Then, for

any $i \in I$, $b \in [\underline{b}, \bar{b}]$,

$$Q_j^i(b) = \frac{1 - Q_0^i(b,p) - Q_i^i(b,p)}{n-1}$$

And,

$$Q_i^i(b,p) \equiv Q_1^1(b,p)$$

Therefore, the expected payment function is the same for all buyers. So buyers i's expected utility is

$$U^i(\mathbf{b}_i, \mathbf{t}_i; \mathbf{p}, y^i) = \langle \mathbf{Q}^i(\mathbf{b}_i; \mathbf{p}), \mathbf{t}_i \rangle - Y^i(\mathbf{b}_i; \mathbf{p}, y^i) \tag{2.31}$$

when,

$$U^i(\mathbf{b}_i, \mathbf{t}_i; \mathbf{p}, y^i) = \langle \mathbf{Q}_i^i(\mathbf{b}_i; \mathbf{p}), \big[\mathbf{v}_i^i - \frac{1}{n-1}\sum_{j \neq i} \mathbf{v}_j^i\big]$$

$$+\big(1 - \mathbf{Q}_0^i(\mathbf{b}_i; \mathbf{p})\big)\big[\frac{1}{n-1}\sum_{j \neq i} \mathbf{v}_j^i\big]\rangle - Y^i(\mathbf{b}_i; \mathbf{p}, y^i)$$

Since b is a symmetric bidding strategy and seller always will sell the object then $\mathbf{Q}_0^i(\mathbf{b}_i; \mathbf{p}) = 0$. Therefore, we have ,

$$U^i(\mathbf{b}_i, \mathbf{t}_i; \mathbf{p}, y^i) = \langle \mathbf{Q}_i^i(\mathbf{b}_i; \mathbf{p}), \big[\mathbf{v}_i^i - \frac{1}{n-1}\sum_{j \neq i} \mathbf{v}_j^i\big] + \big[\frac{1}{n-1}\sum_{j \neq i} \mathbf{v}_j^i\big]\rangle - Y^i(\mathbf{b}_i; \mathbf{p}, y^i) \tag{2.32}$$

Because $\mathbf{Q}_0^i(\mathbf{b}_i; \mathbf{p}) = 0$, and $\mathbf{Q}_j^i(\mathbf{b}_i; \mathbf{p}) = \frac{\left(1 - \mathbf{Q}_i^i(\mathbf{b}_i; \mathbf{p})\right)}{n-1}$ for all $i \neq j$. Buyer's i expected payoff is determined by the bid and the difference between the valuation of the object and the externality the bidder incurs when the other buyer wins the object.

Let b^* be the symmetric bidding strategy with range $b^* \subset [\underline{b}, \bar{b}]^N$ and $\mathbf{b}_i := v^i$ a

where α is negative externality, and $\bar{b} = v_i^i$ where

$$b^*(t_i, t_{-i}) = v_i^i - \frac{1}{n-1} \sum_{j \neq i} v_j^i, t_i \in T \tag{2.33}$$

Equation 2.33 corresponds to the valuation of buyer i minus the average externality from $n-1$ buyers in the auction.

Lemma 1 *When (p, x) is a mechanism by always transferring the object from seller to buyers, then the symmetric bidding strategy b^* is an equiliburium for (p, y). Therefore, for all $i \in I$ and $t_i, s_i \in T$, we have $b_i^*(t_i) = b_i^*(s_i)$ which implies $b_i(t_i) = b_i(s_i)$.*

In symmetric auction, the buyers are not betraying their identities by their bid which is considered as an advantage for such auction. In the following section we have case study analysis to explain the problem with numerical examples.

2.6 Case study Analysis

In this section, we consider the single seller with one object and two buyers who are competing to get the object. Buyers' types have an additional coordinate to represent the externality they suffer when other competitors win the object, and the externality to others if the bidder i win the object. We started with a very specific case with the following model:

$$U_i(E(k)) = \begin{cases} v_j^i - E(k), & \text{if } j = i \\ \end{cases}$$

$E(k)$ is defined as an externality and restores incentive compatibility to force buyers to report the valuation truthfully. $E(k)$ does not depend on the opponents' bids, and as true valuation is paid by the winner in equilibrium. We assumed $E(k)$ function can be written as

$$E(k) = \sum_{j=1}^{k-1} \left(\frac{n}{v_i^i}\right)^{n-1}$$

Given that, $E(k)$ is a strictly convex function of bid k. When we have two different types of buyers, including high-value bidder (e.g., Amazon), and the low-value bidder (e.g., Walmart), and the seller (Whole Foods). Note that, n represents the number of players in the game. For v_i^i of these two bidders when v_1^1 for high value bidder, and v_2^2 for low value bidder are respectively 100 and 50, and a bid of $k = 30$, the negative externality are 4.35, and 8.70 respectively. A bid of 70 is given the negative externality by 24.15, and 48.3, respectively. Finally, the maximal bid of 100 is given the negative externality by 49.5 and 99 for the high-value bidder and low-value bidder, respectively. Figure 2.1 shows the graph-related the growth of externality for both types of bidders based on the amount of bids. The main observation is that the bidder aggressively bids higher to win the auction and avoid the negative externalities regardless of the valuation for the object.

As it is shown in Figure 2.1, the bidder with a higher value for the object will get less negative externality when he bid higher compared to the weaker bidder.

Furthermore, we aim to optimize the welfare of the seller. In the case of $n = 2$,

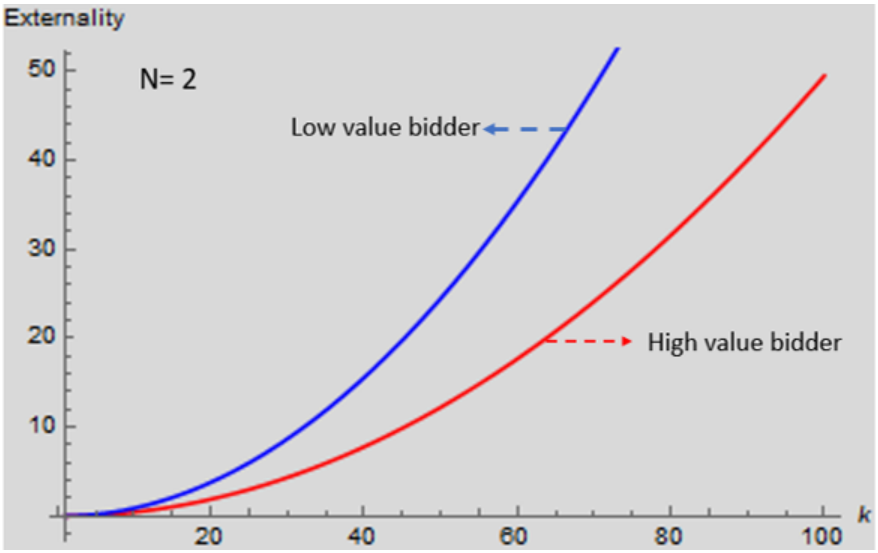

Figure 2.1: Externality function for N=2

by considering the type space for the strong bidder and weak bidder, and α_i as negative externality when bidder i cause to others and β_{-i} as externality when other bidders lose the object if player i wins the object, we have the following matrix.

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \pi_1 & -\alpha_2 + \beta_1 \\ 0 & -\alpha_1 + \beta_2 & \pi_2 \end{pmatrix}. \tag{2.34}$$

With same information for bidder 1, and bidder 2 values (e.g., 100, 50 respectively), and the negative externality of 4.35 for bidder 1, and 8.70 for bidder 2, the maximum welfare for seller would be 71.73 when each bidder has equal probability of winning the auction. When we consider the externality due to β_{-i} to calculate the seller welfare. We could have two different scenarios which are as follows:

(i) $\beta_1 < \beta_2$

When the condition (i) happens, and bidder one as the strong bidder (e.g., Amazon) with higher value lose the object, and bidder two as the weak bidder (e.g., Walmart) wins the object, then optimal utility for the seller, in our case Whole Food is 67. When the condition (ii) happens, and the externality for the strong bidder becomes equal or greater than the externality for bidder 2, if bidder 1 wins the object, then optimal utility for the seller is 69.78. Therefore, the optimal utility for the seller would be higher if the bidder with higher value wins the auction even if the bid is less than the true value. The case study analysis by considering our model shows that the seller's decision toward giving the object away by a higher bid is a good decision, but it is not necessarily maximizing the welfare since the lower value bidder's externality influences the seller's payoff. In particular, if the externality created by a sale reduces the welfare, then the seller is better off by not selling at all. We use numerical studies on Python to compare the outputs of the two scenarios with traditional auction and evaluate the results. Figure 2.2 shows the results related to the scenario 1 where buyer 1 with the higher value wins the auction. Figure 2.3 shows the results related to the scenario 2 when buyer 2 with the lower value wins the auction.

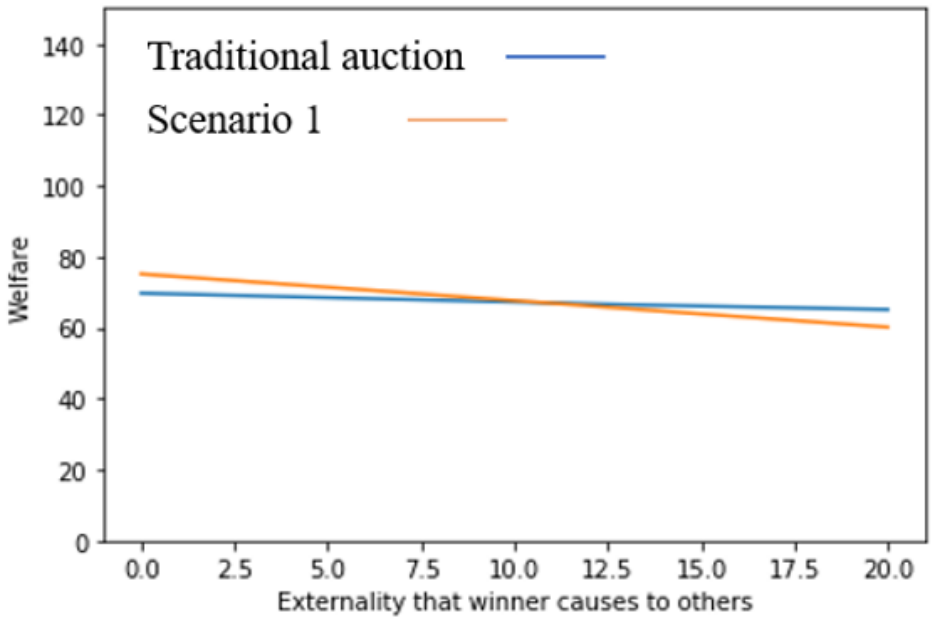

Figure 2.2: Scenario 1 where the buyer 1 with the higher value wins the auction

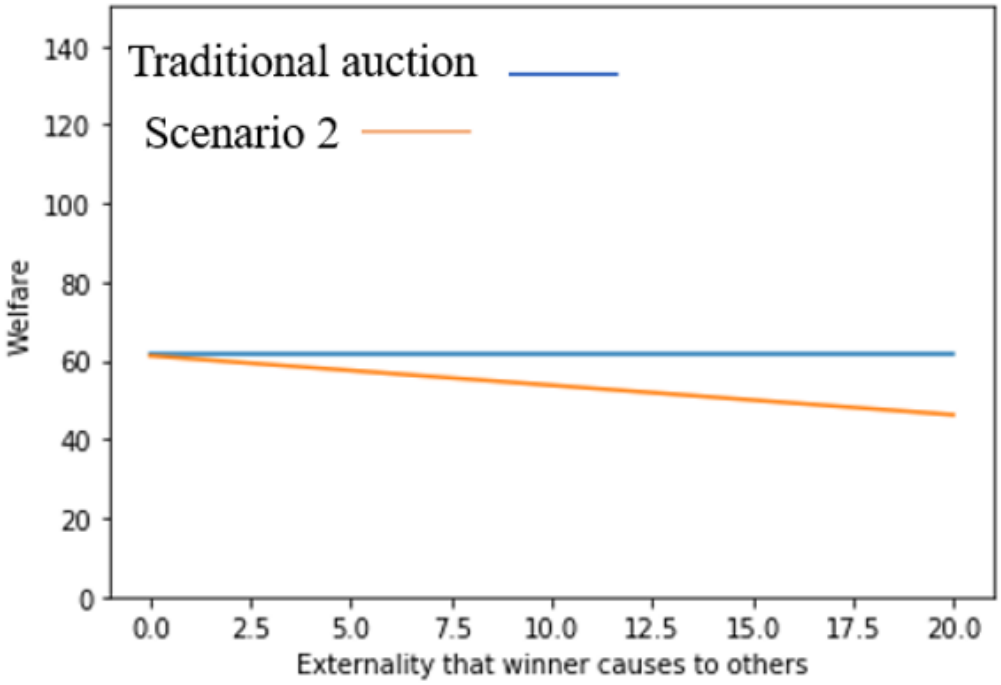

Figure 2.3: Scenario 2 where the buyer 2 with the lower value wins the auction

As it is shown, when the buyer with the higher value on the item wins the auction, the welfare is higher than traditional auction while by increasing the negative externality that winner causes to others, the welfare reduces. In the case of buyer with lower value on item as a winner, the welfare reduces by increasing the negative externality that the winner causes to others. For example, when Amazon and Walmart were competing for acquiring Whole Foods, the analysts said if Walmart makes a bid for Whole Foods, it is less likely to win [6], but they did not mention about the optimal welfare for Whole Foods on the long run. If the sum of externalities in Whole Foods acquisition was greater than the payoff for Whole Foods, then it was not worth it at all for Whole Foods to consider the acquisition.

2.7 Conclusion

The most important challenge in researches related to auction design is the auctioneer knows enough about the true value of buyers for the item, which reveals how much they are willing to pay for the item and how much they are willing to share their true information when they are bidding. With respect to previous studies, our research explains there is a possibility that auctioned object has an impact on the auction's participants who lose the auction, and so as seller's payoff. We simplified the potential mechanism by depriving buyers of their right to absolute non-participation. In our study, the objective is to help the seller identify an optimal mechanism.

[6]"Wal-Mart could enter a bidding war with Amazon over Whole Foods" *CNBC* June22, 2017.

We study the impact of information sharing structure among competing buyers on the seller's revenue with a single object. We explain that the structure of information sharing among buyers and the seller dictate the design of the optimal mechanism, as well as the seller's expected revenue. We find that the seller could benefit from buyers' loss due to competitive relationships among buyers. The competitive relationships between buyers are modeled as a negative externality, if any competitor obtains the object, what it causes to other buyers, and if other buyers gain the object, what is caused by the winner to the buyer.

Moreover, we show that in the optimal mechanism design, the seller can gain the revenue from the bidders who lose the auction. Therefore, maximizing the revenue for the seller is not an efficient mechanism. Our characterization is more tailored towards understanding the type space for bidders, and the information structure of single object auctions with negative externality set up. We have shown that if the negative externalities created by the sale is greater than the seller's payoff, then the seller is better do not auctioning the object away.

In the Whole Foods acquisition case, many small businesses and local delivery companies lost their opportunity in sales, and many big grocery stores like Kroger and Walmart started to change toward investing in online channel infrastructures, which increases the costs for them. Evaluating the pros and cons of the consequence of Whole Foods acquisition from the seller point of views, and the other bidders' point of view is not related to this study, but our results showed that in an incomplete information structure, the negative externalities are strong enough to affect the outcome of mechanism. Our result provides the foundation for various

used auctions. We also believe that this research opens new research ideas for carrying out a similar model in a multi-object auction model. Moreover, the future research can focus on the positive externalities that winner can cause on other bidders and the seller. Implementing the model on big data related to auction can bring several managerial insights on the research as well. In the future research, we can apply the model on real data to evaluate the impact of competitive relationship among buyers on designing the optimal mechanism.

Appendix

2.8 Proofs

Proof of proposition 1: A vector field $v : \mathbf{R^N} \to \mathbf{R^N}$ is conservative if it is a gradient of a function $V : \mathbf{R^N} \to \mathbf{R}$. Hence, if v is differentiable, v is conservative iff $\frac{\partial v_i}{\partial \mathbf{r_j}} \equiv \frac{\partial v_j}{\partial \mathbf{r_i}}$ for all $i \neq j$. That is, iff the Jacobian of v is symmetric. $\square$

Proof of proposition 2: Since Q^i is monotone, S^i is convex. Therefore,

$$S^i(\mathbf{t_i}; \mathbf{p}, \mathbf{y^i}) \geq \mathbf{S^i}(\underline{\mathbf{t_i}}; \mathbf{p}, \mathbf{y^i}) + \langle \mathbf{Q^i}(\mathbf{s_i}; \mathbf{p}), (\mathbf{t_i} - \mathbf{s_i}) \rangle$$

Thus, we have

$$S^i(\underline{\mathbf{t_i}}; \mathbf{p}, \mathbf{y^i}) + \langle \mathbf{Q^i}(\mathbf{s_i}; \mathbf{p}), (\mathbf{t_i} - \mathbf{s_i}) \rangle \geq \mathbf{0}$$

Moreover, we have $\langle \mathbf{Q}^i(\underline{t}_i; \mathbf{p}), (t_i - \underline{t}_i) \rangle \geq 0$, while $\mathbf{Q}^i_i(\underline{t}_i; \mathbf{p}) \geq 0$ and $v^i_i \geq \underline{v}^i_i$. Therefore, $\langle \mathbf{Q}^i(\underline{t}_i; \mathbf{p}), (t_i - \underline{t}_i) \rangle = \sum_{j \neq i} \mathbf{Q}^i_j(\underline{t}_i; \mathbf{p}), (v^i_j - \underline{v}^i_j) \geq 0$

Proof of lemma 1: By symmetry, we show the result for $i = 1$. Surplus function for player 1 is $S_1(t_1) := sup\{U^1(b_1, t_1) | b_1 \in [\underline{b}, \bar{b}]\}^N$.

Let $t^1, s^1 \in T_1$ be such that $b_1(t^1) = b_1(s^1)$. Since $\mathbf{Q}^i_0(\mathbf{b}_i; \mathbf{p}) = 0$ for all $b_1 \in [\underline{\mathrm{b}}, \bar{b}]$, and

$$U^1(b^*_1, t_1) = U^1(b_1, s^1) + \frac{1}{n-1} \sum [v^1_j - s^1_j].$$

Therefore, $S_1(t_1) = S_1(s_1) + (\frac{1}{n-1}) \sum_{j \neq i} [v_j^i - s_j^1]$. Furthermore, S_1 is differentiable at t_1 if and only if S_1 differentiable at s_1.

By assuming all buyers use a common bidding strategy $b(\cdot)$, and $1 < i < j$,

$$\mathbf{Q}^1(\mathbf{b}_1(t_1); \mathbf{p}) = \mathbf{Q}^1(\mathbf{b}_1(s_1); \mathbf{p})$$

, and thus $b_1(t_1) = b_1(s_1)$ by considering $\mathbf{Q^1}$ where,

$$\mathbf{Q}^1(\mathbf{b}_1; \mathbf{p}) \equiv \sum_{j=0}^{n} \mathbb{E}\left[p_2(\mathbf{b}_1, \mathbf{\Theta}_{-i}) \cdot \tilde{\mathbf{a}}_2^1\right],$$

and, $Y^1(\mathbf{b}_1; \mathbf{p}, y_i) \equiv y_i - \sum_{j=0}^{n} \mathbb{E}\left[p_2(\mathbf{b}_1, \mathbf{\Theta}_{-i}) \cdot \tilde{b}_2^1\right]$. $\square$

Chapter 3

Sentiment Analysis on Luxury Products at Amazon

3.1 Introduction

Online shopping offers consumers a wide variety of products with a low search cost, but unlike in the context of brick and mortar retailers (B&MRs), customers cannot try out a product prior to purchasing it online. Thus, they devote considerable time to reading reviews to learn about other customers' experiences. They spend even more time on reading reviews when a product is expensive, and they also find it challenging to learn about the product on the retailer's website. Fashion merchandise presents one of the greatest challenges for online business because customers spend more money in this market and have high expectations for the products. Therefore, selling luxury goods is a very risky business for online retailers (ORs), and recent publications have raised the issue of the compatibility of luxury and the internet (e.g., Hu et al. (2011a, 2012); Salehan and Kim (2016)). Although ORs try to gain consumers' trust by adopting strategies such as having a rating and review system, the continuing presence of fake product reviews reduce this trust. The large body of literature focuses on customizing consumers'

reviews on future purchases (e.g.,Gardner (2011); Tan and Wei (2006); Thongpapanl and Ashraf (2011)). Some of these papers focus on either the effect of product reviews on a product choice or sentiment analysis on positive and negative reviews, and neither of them shed light on evaluating the relevant features for enhancing the performance of classifiers to predict the helpfulness of reviews. Examining numerous online reviews on a company's website enables consumers to make informed decisions about the quality and credibility of products. However, if these reviews are not reliable, they could reduce the sales, especially when a product is expensive. To overcome this problem, Amazon.com added a "verified" tag to reviews written by users who have bought the product [1].

Although this tag may influence the purchase intention and final decision of the user about the product, there exist several fake reviews with "verified" label that are written by sellers who had vouchers to purchase their products and get this label [2]. To solve this problem and regain consumer trust, Amazon.com introduced a "helpful" voting system along with the review text, which allows users to vote on whether the content of the review has valuable information or not. Consumers can then use the number of helpful votes on a review as a reference to make a purchase decision. To reduce the hassle of searching reviews with highest helpful votes, Amazon.com sorts the reviews according to the number of helpful votes on the first page of a product. Several studies have assumed that these votes reflect the quality of reviews' contents. However, the voting system can be manipulated

[1] "Verified Purchase Reviews vs. Unverified Purchase Reviews: Why Does it Matter?" *AMZ advisers* May 16, 2019

[2] "Amazon can't end fake reviews, but its new system might drown them out" *Vox* Feb 14, 2020

by fake user accounts as well. Since consumers usually do not read all the pages of reviews, the reviews with less or none helpful votes will be missed in further pages of reviews. Nakayama and Wan (2017) conducted empirical research and found that fake vote counts can completely change judgments of reviews' quality and influence consumers purchasing decisions. In this study, we shed light on how to predict the helpfulness of reviews and find the most frequent terms in helpful and non-helpful reviews.

To this end, we build a network of terms (e.g., bi-grams, tri-grams) by using the term frequency-inverse document frequency (TF-IDF) technique to analyze the terms' connection in both helpful and non-helpful reviews. In other words, our approach improves the review helpfulness prediction performance by using topics and terms. Since Amazon.com is one of the largest ORs in the world that began selling luxury products by introducing the Amazon Fashion concept to its website, we recognized an opportunity to get access to reviews on this section of Amazon.com.

The main goal of research related to sentiment analysis is to obtain the feelings that consumers express in positive or negative comments that garner more helpful votes and predict the helpfulness of reviews. The research questions for this study are as follows (i) Can we predict the helpfulness of a review? (ii) What are the possible reasons of getting more helpful votes? and (iii) What are the most frequently repeated keywords in helpful and non-helpful reviews?

To answer the research questions, we demonstrate our model's performance in prediction accuracy by extracting the main features from topics analysis using

the latent Dirichlet allocation (LDA) model, and topics analysis plus bi-grams by using the TF-IDF technique. We then use these features in a support vector machine (SVM) classifier by having a binary response variable, namely helpful and non-helpful reviews. We compare the performance of two models : (i) topics analysis using LDA features for SVM classifier, and (ii) topics plus bi-grams with TF-IDF vectorizer for SVM classifier. The techniques operate on a large corpus of Amazon Fashion review texts and predict the helpfulness of reviews by having a large data set related to product information.

We examine the performance of classifier by using the binary response variable to assess the number of helpful votes on reviews. We consider reviews with more than three helpful votes as helpful reviews and otherwise as non-helpful reviews. We use the available dataset of Amazon Fashion reviews, we perform classification experiments on samples from different product categories for each iteration. In comparison with previous methods, our method tends to be more accurate since we train the classifier by using real-world dataset. Furthermore, we use a test set to determine the accuracy of the system. We prove the effectiveness and capability of the topics plus bi-grams TF-IDF vectorizer model by comparing the performance of both approaches, namely topics (LDA)-SVM, and topics plus bi-grams TF-IDF vectorizer-SVM. To validate the models, we use a cross-validation technique and split the training set into five-folds.

Moreover, we use a receiver operating characteristics (ROC) curve to evaluate the performance of each model with cross-validation. We also examine the possible reasons why reviews gain more helpful votes by performing an exploratory analysis

on terms and their correlations with helpful votes. We find that the longer reviews do not obtain helpful votes, and consumers most likely choose to read the reviews that are neither too short nor too long. We also find that the star rating system is useful at the primary stage of the purchasing process, but it does not attract consumers' attention when the product is expensive, and thus the consumers likely rely on reading the reviews. Furthermore, we find that in the fashion industry, consumers are sensitive to the appearance and the quality of products rather than their price.

Therefore, the most frequent positive terms on helpful reviews are stylish, soft, handy and light, showing that customers in the fashion filed care about the style and material of clothes because they seek to establish a personality signature and feel comfortable, respectively. The most frequent negative terms are bent, return, junk, and stiff which show that consumers' are concerned about the fabric and the return experience. In sum, in this study, we build a classifier model that can predict the helpfulness of reviews based on the content of a review and identify its helpfulness irrespective of the volume of votes.

We organized this study by reviewing the literature in §3.2, and particularly reviewing the background of using text mining in the fashion industry. In §3.3, we introduced the data set by having a overview on the reviews in subsection §3.3.1, and a overview of the review helpfulness in subsection §3.3.2. In §3.4, we have data process which includes §3.4.1, §3.4.2, §3.4.3. In §3.5 we select the appropriate features for our study which are introduced in §3.5.1 and §3.5.2. We introduced supervised machine learning algorithm, namely support vector machine (SVM) in

§3.6, and includes subsections §3.6.1, §3.6.2, and §3.6.3. The results of the study is in §3.7 which includes subsection 3.7.1, and §3.7.2. Finally, we have the conclusion and future work at §3.8.

3.2 Literature Review

Online reviews on online channels have become the main source of information for consumers to learn about products before purchasing them. Iyengar et al. (2011) found that reviews on websites impact consumers' decisions about which product to buy. Our research contributes to four research streams namely, sentiment analysis and manipulated reviews, helpfulness of reviews, and application of text mining in the fashion industry. A large body of literature has accumulated recently on the work involved in sentiment analysis of texts containing personal opinions. Pang and Lee (2004); Pang et al. (2008, 2002) used several machine learning systems to implement a binary classification task of movie reviews to examine users' opinion on different genre of movies.

Several scholars (e.g., Pan and LIN (2008); Salehan and Kim (2016); Salvetti et al. (2006)) used structured reviews for testing and training for determining the polarity of reviews. Some researchers applied the sentiment classification strategy based on supervised machine learning classification methods, including naïve bayes, SVM, bayesian network, decision tree, and random forest algorithms for sentiment classification of Twitter data for several services, including airline service (e.g., Catal and Nangir (2017); Kanakaraj and Guddeti (2015); Shrivastava and Nair (2015); Wan and Gao (2015).

The second research stream is related to reviews manipulation when consumer's reviews and their vote on the product or helpfulness of the reviews are manipulated to enhance the sales of specific products. Several studies have investigated and confirmed the presence of manipulated reviews on online review platforms for several services (e.g.,Hu et al. (2011b); Ott et al. (2011); Sharma and Lin (2013)). However, these studies have primarily focused on the rating system to investigate the presence of online review manipulation without considering the content of reviews. Some studies have considered the content of reviews and proposed methods for identifying products with manipulated reviews (e.g.,Hu et al. (2012); Luca and Zervas (2016); Ludwig et al. (2013)). Furthermore, some scholars focused on determining fake reviewer groups by identifying suspicious patterns in the contents of reviews (e.g.,Kolhe et al. (2014); Wang et al. (2015)). We enhanced the performance of our models by regretting those manipulated reviews.

The third research stream is related to review helpfulness and its impact on consumers' future purchases. The literature in this area has divided into two sections. The first section concerns the effects of review contents on the helpfulness of reviews (e.g.,Connors et al. (2011); Korfiatis et al. (2012); Mudambi and Schuff (2010)). The second section of the literature emphasizes on the possible reasons that make the reviews more helpful for consumers (e.g., Cao et al. (2011); Huang et al. (2015); Willemsen et al. (2011)). Korfiatis et al. (2012) explained as part of their results that there is a positive correlation between helpful votes and the review length. Furthermore, they explained that the helpful reviews are usually positive or strongly positive with the longer explanation about the product.

Mudambi and Schuff (2010) developed and tested a model of customer review helpfulness by analyzing 1,587 reviews from Amazon.com across six different products and found that the correlation between the star rating system and helpful votes are weak. In contrast, review depth (extremely positive or extremely negative) has a positive effect on the review's helpfulness. Furthermore, they explained that lengthier reviews generally increase the helpfulness of the review. We test this assumption in our study to examine if the length of the review has a positive correlation with its helpfulness in the fashion field.

Finally, the fourth research stream is related to the importance of the text mining approach in the fashion industry. Several studies proposed text mining method on review data to predict color or style trends in the fashion industry field(e.g., An and Park (2017); Romão et al. (2019)). Dennison and Montecchi (2017) explained the effect of online consumer reviews on female fashion consumers in terms of subsequent purchase decisions. Their results showed that reviews with credibility and positive words significantly increased the purchase intention of female fashion consumers. To the best of our knowledge, there is no research related to using a text mining approach on consumers' reviews for fashion products on the online channel. The reason might be the lack of enough data related to this field. Given that Amazon released the data related to Amazon fashion a few months ago. Our research fills the gap by using a supervised machine learning algorithm to predict the helpfulness of the reviews and seek to find the most common words that are used in a helpful review in the fashion field.

3.3 Dataset

To implement our study, we focused on a subset of Amazon.com product review data related to the Amazon Fashion category [3]. In particular, we are using the dataset for the sensory products like fashion products and beauty products sold through Amazon, which were almost 80,000 reviews on Amazon Fashion with 18,637 products spanning May 2004 - July 2018. We clean the data and narrowed it down to 79,611 reviews after cleaning. The data has two different sets of information. One set of information is related to consumers' reviews, and the other set of information is related to the products' descriptions. The purpose of this research suits to the first category of dataset which is related to consumers' reviews.

Consumers' reviews have the product ID, title of the product, product's price, user ID and the name of the reviewer, the number of users who found the review helpful, the star rating for the product (out of five), verified review, time of the review, review summary, the text body of the review, and the product description. Product information includes category information, price, brand, also viewed, also bought. The following table shows all the review attributes along with their descriptions. In the next sections, we have an overview of reviews and the review helpfulness, respectively.

Table 3.1: Attributes Descriptions

Review Attribute	**Description**
Reviewer ID	Unique identifier for the user
Asin	Unique identifier for the product
Reviewer Name	User profile
Vote	Helpful votes of the review
Review Text	Text of the review
Overall	Rating for the product
Summary	Summary of the review
Unix Review Time	Time of the review (Unix time)
Review Time	Time of the review

3.3.1 Overview of the Reviews

Each review includes a review text which is a detailed description of the product from users' perspective, and their opinion about that product. Some of the examples of the review text are shown below:

- *I was looking for a tab collar dress shirt. I have always liked that style, but lately, I had been unable to find any. I looked on Amazon and lo and beheld; there they were. Reasonable price, fast delivery, and excellent quality. I have absolutely no complaints, and I intend to purchase several more in the near future.*

- *Perfect out of the box. I have worn a lot of dress "business" over the years, and I know these will be a favorite. Great style. Great fit. Great value.*

- *The skirt was stuck together when I received it and when I pulled it apart, the color was removed from some of the fabric. The bodice of the dress is stiff and very itchy. It is blatantly cheap, and the cape has a very small neck hole. I will be returning this item. I included pictures of the front and back.*

Furthermore, there is a summary text which conveys information in a few words. Some of the examples of them are shown below:

- *Terrible and cheap*
- *Cute but disappointing*
- *Five Stars!*
- *a good and comfortable costume*

In this research, we extract the feature from both the review text, and the summary text by using n-grams and topics analysis to enhance the accuracy of prediction for our classifier.

3.3.2 Overview of the review helpfulness

The helpfulness of a review is measured by the volume of "helpful vote", which denotes the total number of users who found the review helpful. In this study, we categorize the reviews to helpful and non-helpful reviews where the helpful reviews have higher than three votes.

3.4 Data Process

In this section, we discuss the process of building our text classification system to predict whether or not the reviews of luxury products on Amazon.com are helpful. The process includes the following steps:

(i) Implementing exploratory data analysis to generate the binary response

(ii) Performing text retrieval and text relevant to clean the documents.

(iii) Performing the approaches related to sentiment classification including topics analysis by using LDA methods and topic analysis plus bi-grams by using TF-IDF methods to extract the features.

(iv) Applying the features on support vector machine (SVM) classifier to predict the helpfulness of reviews.

(v) Evaluating the model validation and the accuracy of approaches by using confusion matrix and k-folds cross-validation.

Figure 3.1 shows the process of our research on helpfulness of reviews briefly.

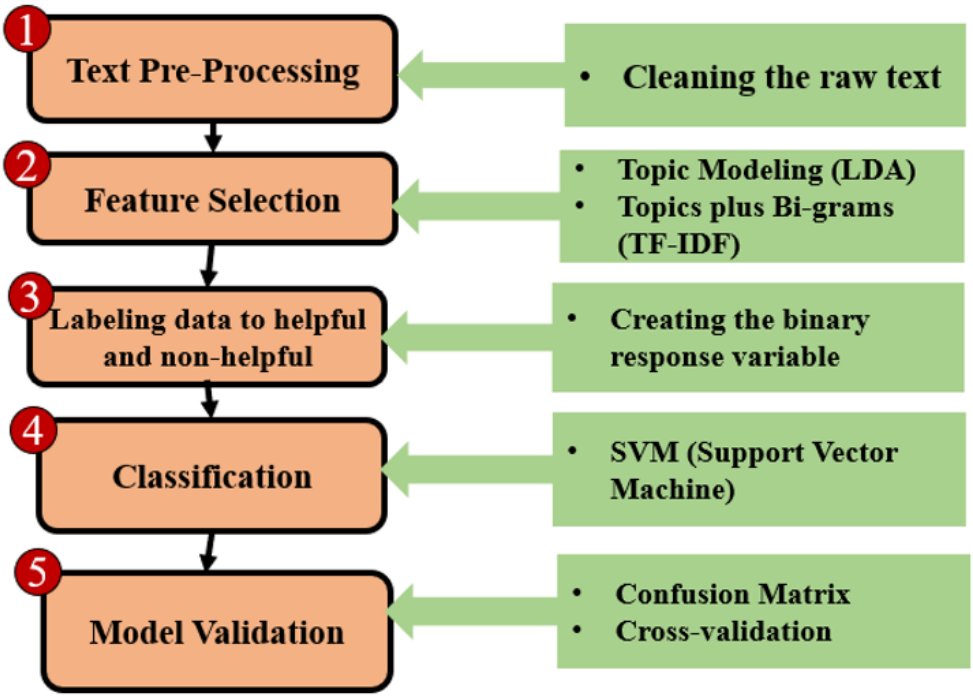

Figure 3.1: Graphical process for text-analysis on Amazon Reviews

3.4.1 Exploratory Data Analysis

This section describes the process of binary response variable generation. By considering the goal of this research, as we mentioned earlier, the non-helpful reviews have less than three votes. In Figure 3.2, we have the frequency distribution of the helpful reviews for the data set where 1 refers to helpful reviews and -1 refers to non-helpful reviews.

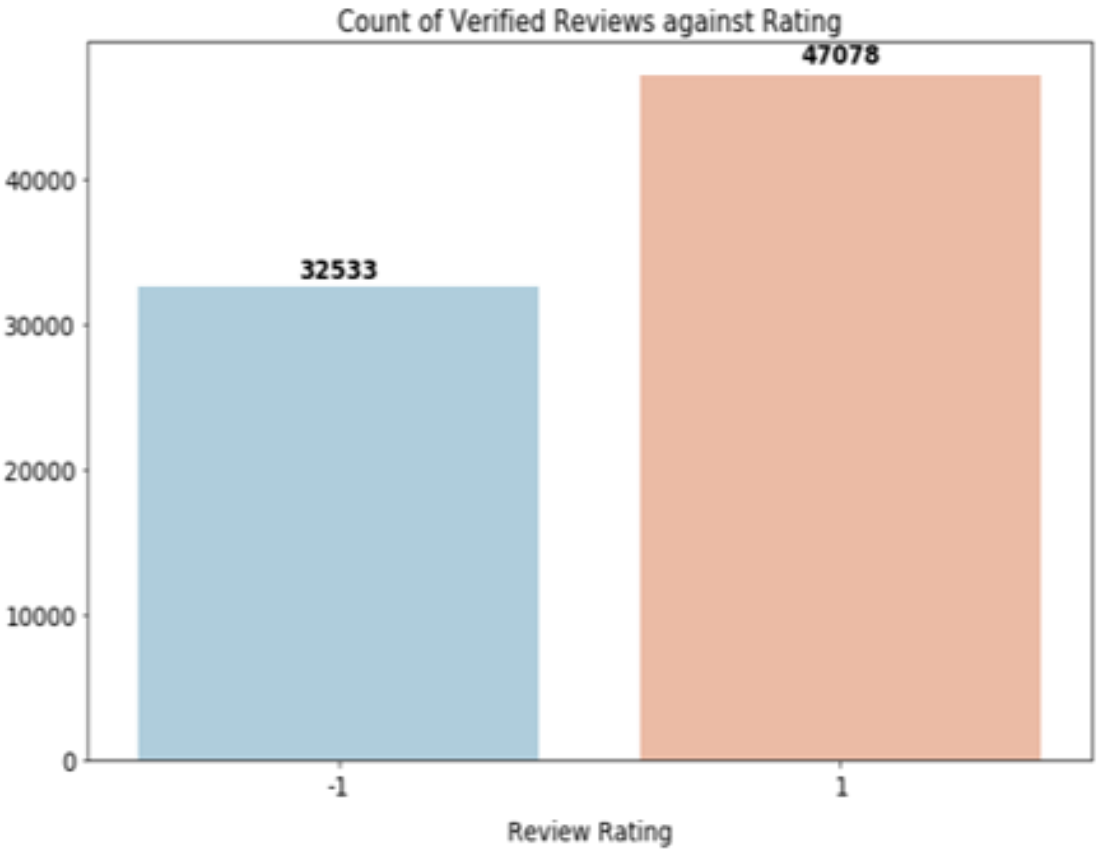

Figure 3.2: Frequency distribution of helpful verified reviews

As Figure 3.2 shows, the helpfulness attribute has two parts, and the number of reviews with more than three helpful votes is larger than the number of reviews with less than three helpful votes. We want to have a fair number of distribution for both helpful and non-helpful reviews, and thus we choose median number of

distribution of the rating system attribute for the helpful reviews. The rating system consists of the integer values from 1 to 5 with "1" as "extremely low" and "5" as "extremely high". Figure 3.3 shows the distribution of the rating system attribute. As Figure 3.3 shows, the distribution of scores related to rating attribute

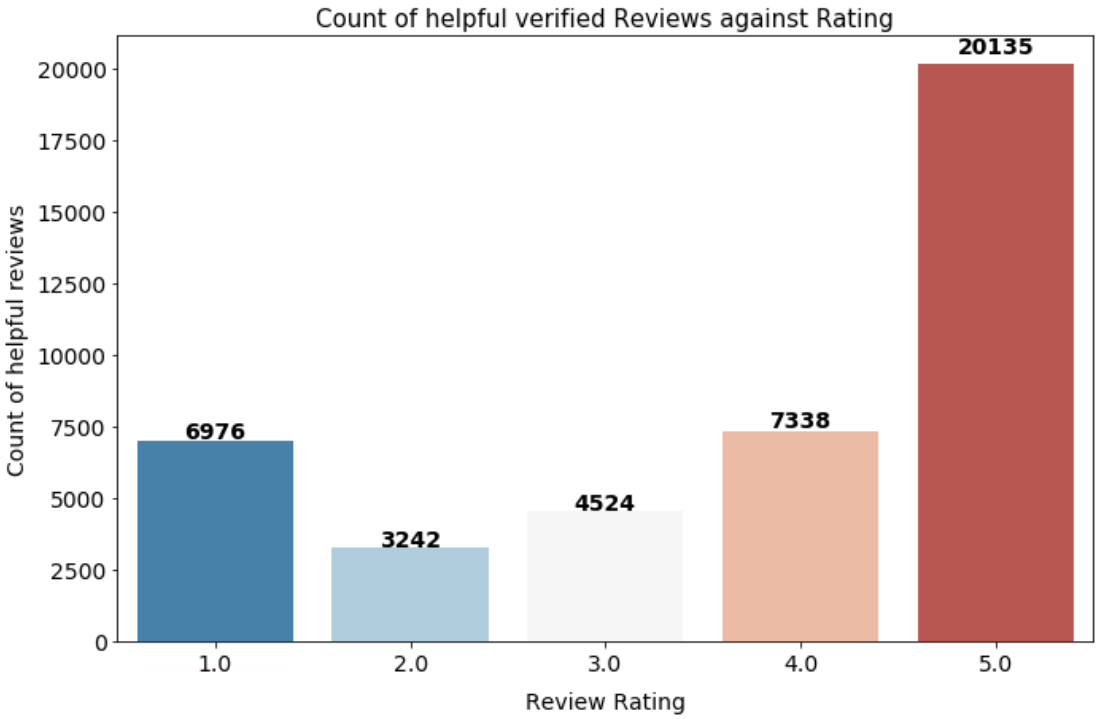

Figure 3.3: Distribution of the rating score attribute for helpful verified reviews

is highly skewed toward the right and the score "5", which shows the majority of reviews with at least three helpful votes have "extremely high" rating. Note that some of the reviews did not have star-rating, and consumers only rely on writing their reviews without giving any star rating to the products.

Furthermore, Figure 3.4 shows the distribution of reviews length. The distribution shows a large number of reviews have between 100 to 200 characters, and few reviews are considered as long reviews with more than a thousand characters. We also have the distribution of length of helpful review in Figure 3.5 which shows

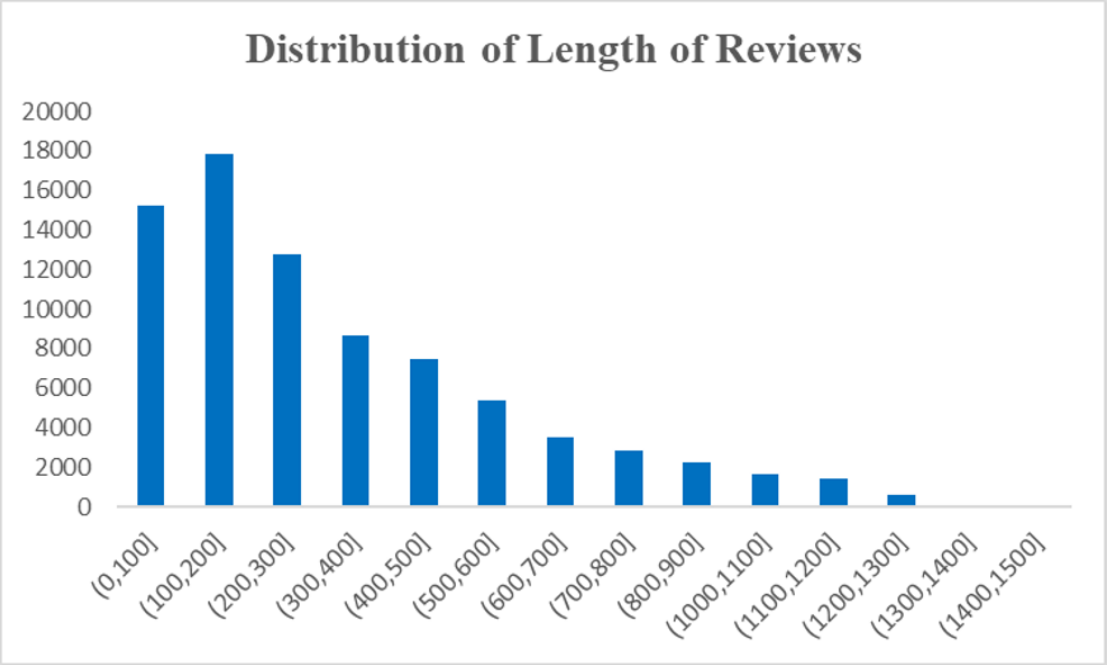

Figure 3.4: Distribution of the reviews' length

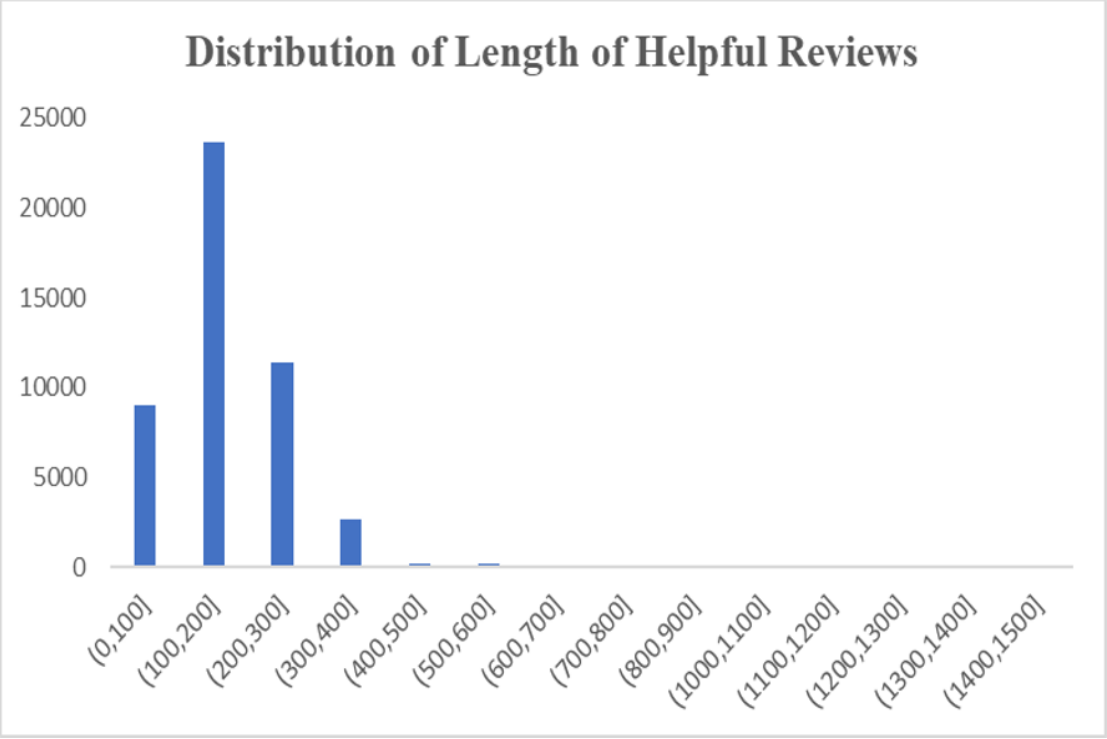

Figure 3.5: Distribution of the helpful reviews' length

Figure 3.6 shows that the distribution of reviews' helpfulness versus the length of reviews is highly skewed toward shorter length reviews. In other words, the

review that is not too short or too long is getting more votes as being a helpful review. Therefore, the possible reason of getting more helpful votes for longer review is not correct in the fashion field. Thus, the lengthier reviews are not necessarily, the more helpful ones on the fashion field.

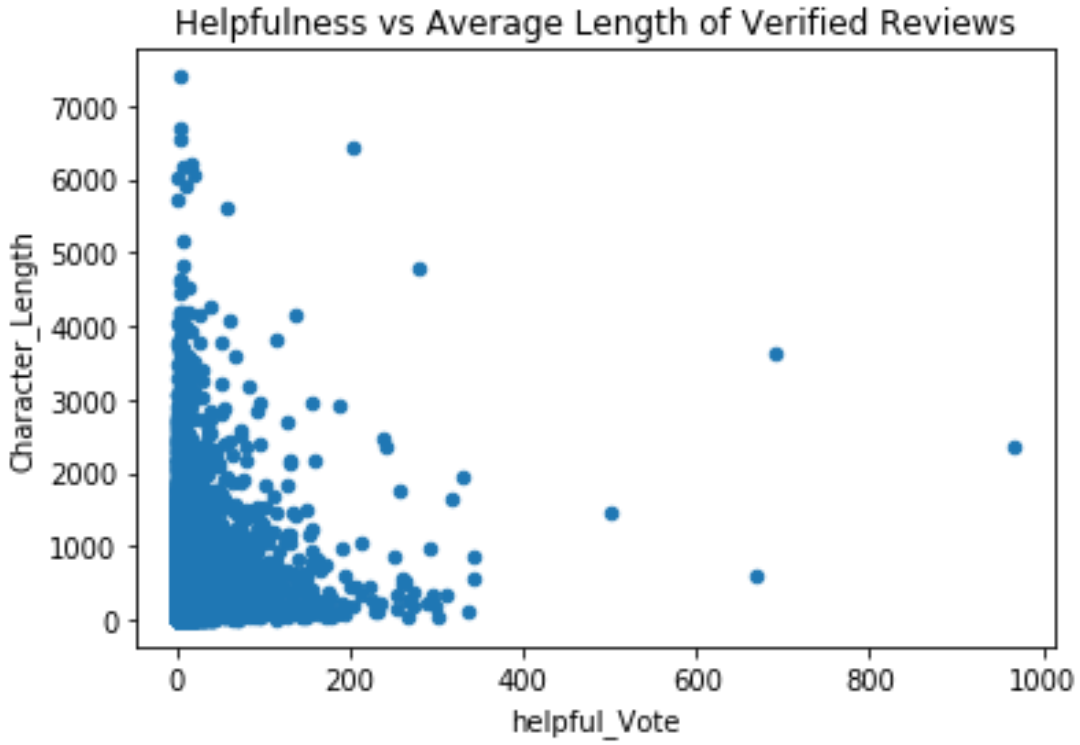

Figure 3.6: Distribution of helpfulness vs length of verified reviews

As mentioned earlier, most reviewers give a five-star rating to helpful reviews and very few give two or three stars. Thus, we conclude that the average rating contains four or five stars, but we investigate how effective the rating system is for selling purposes when the consumer is going to purchase a product with a high star rating. Is it helpful for consumers to learn about a product when they see that it has a higher star rating (e.g., 4.3 out of 5 stars for Levi's Women's Wedgie Skinny Jeans)? As Figure 3.7 shows, the rating system has a very weak correlation with the helpfulness of reviews. Therefore, the correlation is not strong enough to conclude that a higher star rate could yield more helpful votes for reviews. Thus, the possible reason of gaining more helpful votes for the higher star rating (e.g., four stars, five stars) is not clear in the fashion field.

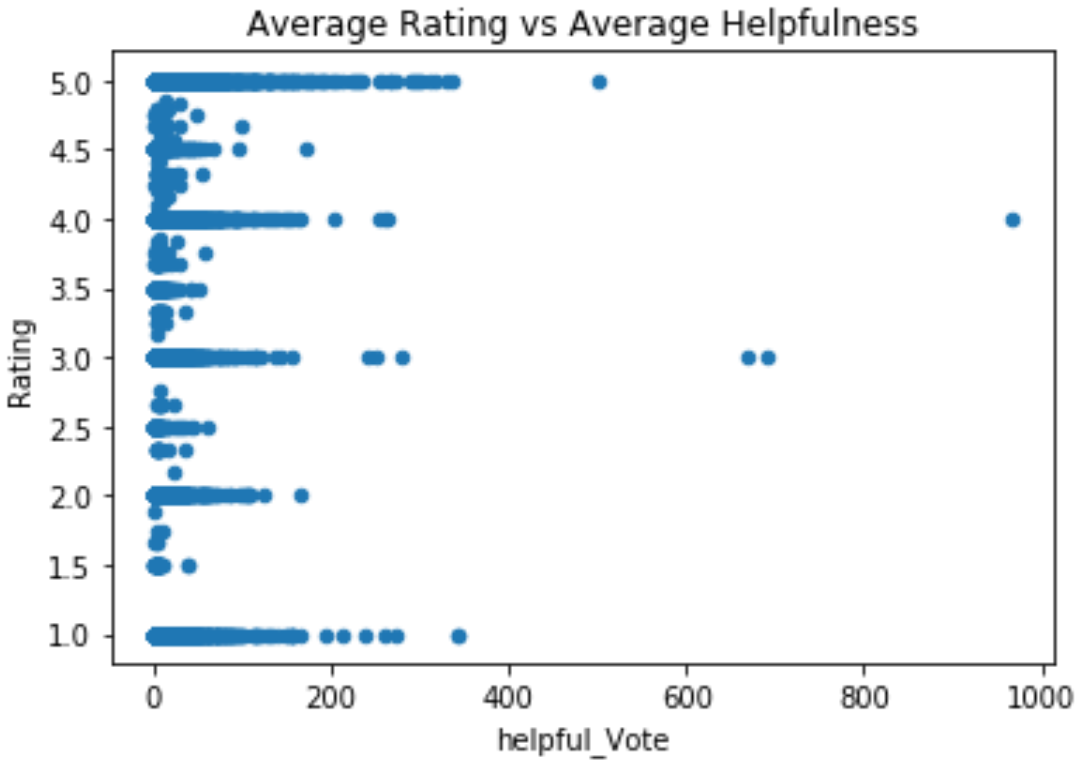

Figure 3.7: Distribution of helpfulness vs stars rating system

3.4.2 Text Preprocessing

Before drawing features from the dataset, we did preprocessing on raw texts. We iterated through over 80,000 reviews for fashion products and narrowed down our dataset to 79,611 unique reviews after cleaning the dataset. Since the raw texts have slang, emojis, and unstructured texts, the data gets clean and reformat with better quality, which has a significant impact on the performance of algorithms. In this study, we did text preprocessing by following components such as string cleaning, tokenization, stop-words removal, lemmatization, and stemming:

- **String cleaning**: cleaning texts from unwanted or useless characters that do not contribute any fundamental meaning to the reviews, such as email address, image link and regular expression by adding emojis. (e.g., :) , and :(and so on).

- **Tokenization**: by having a sequence of texts, tokenization is a task of breaking a chain of textual content into words or phrases.

- **Stop-word removal**: this task is for removing the words that do not have a meaningful content in the document (e.g., “the” , “a”, “an” , “in”, “at”). Three types of words are considered as stop-words : (i) Prepositions, conjunctions, pronouns. (ii) Words are repeating frequently in all the documents without adding information to a review (e.g., “is”, “are”, “want”, “buy”). (iii) Words are appearing few times in the document, which represents no significant information (e.g., “hate”, “punctuation”).

accuracy and efficiency of the analysis by reducing the variation of words within a document.(e.g., “buy”, “bought”, “view”, “viewed”, “dress”, “dresses”)

- **Stemming**: this task is similar to lemmatization, while the method is more straight-forward.

3.4.3 Text Retrieval and Text Relevant

We use text retrieval technique in order to search the relevant text. This technique can prevent the fake reviews that are irrelevant to the products within the reviews. McMahon et al. (2004) explained document retrieval can facilitate the classification tasks. In this study, we generate structured representations of documents by converting the unstructured text of each review into a numeric vector. Particularly, with a pool of documents $D = \{d_1, d_2, ..., d_n\}$, and $V = \{w_1, w_2, ..., w_n\}$ is the set of vocabulary contained in this pool. Considering $f^d(w)$ as the weight of the term w for document d, we have binary values, identifying whether term w is in document d with 1 and 0 is not in document. Therefore, d represented a vector with binary values as follows:

$$d = [f^d(w_1), f^d(w_2), ..., f^d(w_n)] \tag{3.1}$$

This is the naive version of vector space model by having binary values, while we use term frequency-inverse document frequency (TF-IDF) weighting that was

introduced by Salton and Buckley (1988), and the weight is calculated as :

$$f^d(w_i) = \frac{f^d_{w_i}}{\sum_{k=1}^{v} f^d_{w_k}} . log \frac{n}{F_{w_i}} \tag{3.2}$$

Where $f^d_{w_i}$ is the occurrence frequency of term w_i in document d, v is the size of vocabulary V, n is the total number of the documents in the documents, and F_{w_i} is the number of document which has term w_i. By using TF-IDF configurations and a weight threshold, we could extract the most informative keywords and labels for each review. In other words, if a term appears frequently within a review, then the term is important for the review's content. Furthermore, we analyze the polarity of the reviews by creating the word cloud for both positive and negative reviews. Word cloud represents the more frequent words that are appearing on the dataset. We import the libraries including numpy, pandas, matplotlib, collections and wordcloud in Python to create the word cloud for our texts. Furthermore, we use valence aware dictionary and sentiment reasoner (VADER) tool to label the semantic of texts based on their polarities. Therefore, the words clouds are showing both frequent positive words and frequent negative words. Figure 3.8 shows the words cloud related to positive reviews.

Figure 3.9 also shows the words cloud related to negative reviews. Figure 3.8, and Figure 3.9 show the pattern of positive and negative words within the raw texts. In the future sections, we reach to the list of more specified positive and negative words.

Furthermore, we examine the distribution of positive, neutral, and negative

Figure 3.8: Words cloud for positive reviews

Figure 3.9: words cloud for negative reviews

reviews that are categorized as helpful and non-helpful reviews in Figure 3.10. As Figure 3.10 shows, the high percentages of reviews are positive while there is not a significant difference between helpful and non-helpful reviews.

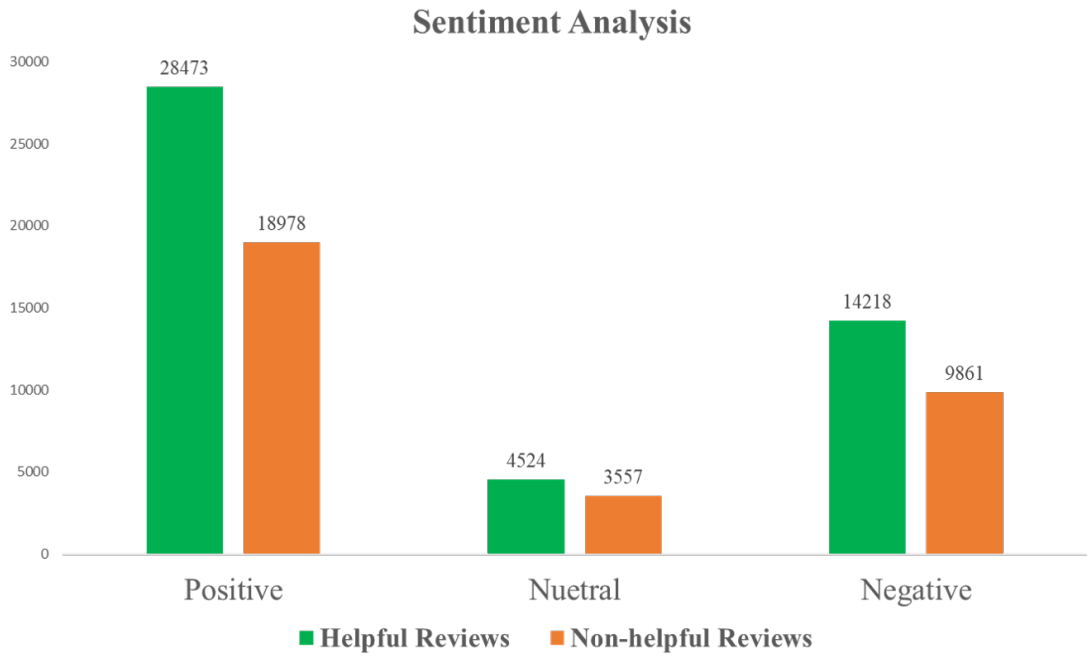

Figure 3.10: The distribution of the reviews' polarities based on their helpfulness

3.5 Features Selection

The other common task in text mining is text classification, which is very applicable in diverse fields. Mitchell (1997) explained that text classification emphasizes to assign the predefined classes or labels to the documents. Singh et al. (2013) did a sentimental analysis on the movie review, which focused on the users' feedback and predicted their interest to recommend movies to them. One of the popular examples of using sentiment analysis related to text classification is customer reviews on the products to predict the product feature in the future (e.g.,Jiang et al. (2017)). The main task in sentiment analysis of reviews is "text polarity classification", where the documents are determined as positive and negative. In our case, we examine the helpfulness of reviews regardless of the reviews' polarity. According to literature, there are two approaches to this problem. One approach uses a sentiment lexicon by having a list of words with

known sentiment helpfulness, which shows the helpful words about the product or the non-helpful words. To the best of our knowledge, there is not such a sentiment lexicon list for the helpfulness of reviews. Note that, there exists a list of words based on their polarity when positive words are associated with the rate of five stars, and negative words about the product are associated with the rate of one star or two stars. Using this approach for analysis of big data is difficult due to presence of the noise in the data. Furthermore, we do not have the list for the helpfulness of reviews.

The second way is building a "model" of the language used for helpfulness of reviews using training data. Machine learning methods involve training of models on documents by adopting supervised algorithms to extract the features related to the data. We first select the "features". The features of sentiment analysis have three dimensions. First, we should examine the basic units extracted from texts, including words and phrases. Second, we need to explore feature selection by considering the frequency of those words and phrases. Third, we should explore feature generalization. Previous research (e.g., El-Din (2016); Martineau and Finin (2009); Wang et al. (2014)) used a standard "bag of words" which has vector of words within the document and is still widely used today. The vector produced by the vector space model (VSM) or bag of word (BOW) model has several elements, and each element indicates the TF, TF-IDF, or appearance of a particular item. We examine the frequency of top twenty words in helpful and non-helpful reviews and extract the top ten positive and negative words by using BOW method. Figure 3.11 shows the most frequent words in terms of helpful and non-helpful reviews.

Noun			
		Helpful	**Non-Helpful**
Rank	Word	Frequency	Frequency
1	Like	15229	8109
2	Great	9311	5685
3	Love	9265	5668
4	Well	6858	3848
5	Good	6651	3979
6	Nice	5888	3724
7	Perfect	5375	2996
8	Top	4483	2140
9	Cute	3936	2654
10	Comfortable	3684	2191
11	Cheap	2614	1524
12	Return	1854	1079
13	disappointed	1742	1021
14	Terrible	1466	1684
15	Worn	1353	780
16	Problem	1281	644
17	Bad	1227	694
18	Issue	769	377
19	Expensive	753	364
20	Wrong	699	430

Figure 3.11: Most frequent words based on helpfulness by using BOW technique

As it is shown in the table, “Like” is a very general word and one the most frequent words in both helpful and non-helpful reviews with the frequency of 15,229 and 8,109 respectively. Furthermore, “Terrible” is highlighted as one of the frequent words in non-helpful reviews which is very general adjective for describing a product. In contrast, “Worn” has the highest frequency for the negative helpful reviews. In sum, the BOW technique is not a efficient technique for this research to

identify the helpful and non-helpful reviews. Therefore, we use topics analysis by using Latent Dirichlet Allocation (LDA) technique to get more particular features for enhancing the performance of our classifier. In the following section, we explain the topics analysis by using LDA technique as one of the techniques that is used to extract the features from our data.

3.5.1 Topics Analysis by Using Latent Dirichlet Allocation (LDA) model

Latent dirichlet allocation (LDA) is considered as unsupervised machine learning technique that examine a document as a limited number of topics. In addition, each topic is a mixture of a number of words. We aim to use this technique for extracting the features to use in our future analysis and to predict the helpfulness of the reviews with the higher accuracy. Blei et al. (2003) introduced LDA model as a more complete generative model. For instance, instead of classifying the sentiment topic by indicating the polarity of movie reviews (e.g.,Pang and Lee (2004)), the researcher can recover the topics based on the genre of movies. In LDA model, we observe the words "**w**" and documents "d". We have $\mathbf{w} = w_1 \dots w_N$ that consider all the documents contain N words in total. Therefore, a document "d" is a vector of N_d words. In this model, we have the matrix of topic distribution which denoted as ϕ with a multinominal distribution over "V" vocabulary for "T" topics that are individually extract from Dirichlet (β) prior.

The matrix of document-specific mixture is θ for "T" topics, and each topic is individually drawn from a symmetric Dirichlet (α) prior. For each word, we have

"z" as the topic that generates that word, extracted from the θ distribution for the document and the word (i.e., w_1) drawn from the topic distribution ϕ that corresponds to "z". Figure 3.12 shows the generative topic model for LDA.

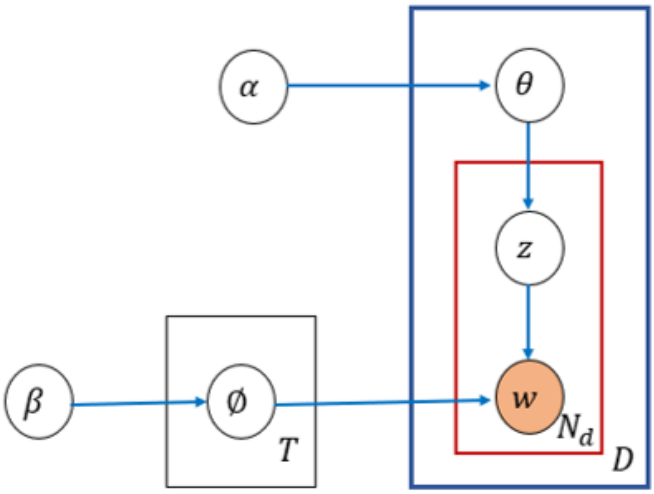

Figure 3.12: Latent Dirichlet Allocation (LDA), a topic model

We construct the multiple LDA models for 15, 20 and 25 topics with learning decay of 0.5, 0.7 and 0.9 to examine all the possible combination of model perplexity. The best number of topics for LDA model in our dataset is chosen by having 15 topics. Each topic shows the words with highest frequency along with document topic weights. The reason we use higher number of topics for this dataset is the reviews are the long texts with no limitation on the amount of characters, and thus we have a lot of new words that are not repeated frequently. The Figure 3.13 shows the weight of each topic in each document, and the dominant topic in each document.

For instance, the most frequent words in topic 1 and topic 2 are as follows:

Topic 1: 'adjustable', 'clean', 'worked', 'solid', 'worn', 'impressed', 'shiny', 'strong',

	Topic0	Topic1	Topic2	Topic3	…	Topic15	dominant_topic
Doc0	0.02	0.02	0.02	0.02	…	**0.45**	**15**
Doc1	0.02	0.02	0.02	0.02	…	0.02	**4**
Doc2	0	**0.13**	0.06	0	…	0.06	**7**
Doc3	0.02	0.02	0.02	0.02	…	**0.35**	**10**
Doc4	0.01	0.01	0.01	0.01	…	0.01	**11**
…	…	…	…	…	…	…	…
Doc79,611	0.02	0.02	0	0.02	…	**0.32**	**15**

Figure 3.13: Document Topic Weights

'pleased', 'best', 'expensive', 'fake', 'fast', 'perfectly', 'right', 'wonderful', 'happy', 'sturdy', 'works', 'beautiful', 'easy', 'awesome', 'excellent', 'comfortable', 'amazing', 'work', 'gold', 'perfect', 'recommend', 'great'

Topic 2: 'clear', 'fun', 'bad', 'loose', 'uncomfortable', 'sad', 'perfect', 'lost', 'bright', 'tops', 'disappointing', 'cheaply', 'hard', 'worked', 'dark', 'perfectly', 'charm', 'right', 'excited', 'good', 'fell', 'accurate', 'comfy', 'fine', 'lovely', 'adorable', 'disappointed', 'super', 'pretty', 'cute'

Therefore, we examine the most frequent adjectives and verbs on each topic based on their helpfulness (i.e., helpful and non-helpful). Figure 3.14 shows top three frequent words in each topic for the helpful and non-helpful reviews.

	Topic Analysis- LDA			
	Helpful		Non-Helpful	
Topics	Adjective	Verb	Adjective	Verb
1	comfortable, worth, sexy	return, recommend, refund	difficult, hard, unfortunately	loved, like, amazed
2	cheap, flimsy, waste	loose, refund, excited	bother, easy, disappointing	fell, work, loves
3	durable, issue, worth	support, satisfied, broke	poorly, free, problem	satisfied, excited, hated
4	charm, problem, weird	adore, loved, smelled	horrible, fake, bad	disappointed, return, hate
5	skinny, cute, wrinkle	worked, refund, falling	worn, bust, nicely	loved, ripped, liked
6	comfy, warm, soft	wrinkle, worn, recommend	decent, ripped, hang	losing, works, broke
7	helpful, stunning, remarkable	exciting, trilled, amazed	terrible, stiff, solid	loose, returned, exciting
8	return, right, issue	returned, smells, refund	expensive, loved, excellent	bought, hate, worn
9	affordable, classy, adjustable	worked, falling, broken	best, happy, pleased	stuck, broken, worked
10	favorite, thicker, uncomfortable	secure, helped, loved	easy, unbiased, fun	amazed, like, lost
11	fake, thinner, solid	Complain, hate, disappointed	enough, better, fantastic	flattering, break, buckle
12	right, flattering, odd	fell, broke, excited	horrible, fun, nicer	hating, thrilled, disappointed
13	odor, complaints, sturdy	lost, thank, disgust	Weird, junk, glad	lost, recommend, sting
14	gorgeous, sharp, problem	adjust, sparkle, amazed	Funny, cheaper, best	works, fell, amazing
15	poor, beautiful, correct	support, fell, loves	Horrible, nice, good	hate, return, refund

Figure 3.14: Topic Analysis based on helpfulness

The verbs "loved, likes, amazed" have the highest frequency in non-helpful reviews while in helpful reviews they have either lower rank or they do not exists on top three frequent terms. The verbs "return, recommend, refund" have the highest frequency in helpful reviews which shows consumers rely on the reviews on their future purchase if they find these terms on the reviews. We extract the features from topic analysis to use for supervised machine learning algorithm. In addition, we compare the results of a supervised machine learning algorithm by using only topics analysis (LDA) features, and by using both topics analysis plus bi-grams with TF-IDF vectorizer to evaluate which one has a better performance for predicting the helpfulness of reviews. In the following section, we explain the n-grams analysis to extract the features for our classifier.

3.5.2 N-grams Analysis by Using TF-IDF Vectorizer

Since fashion products are expensive and they have their target consumers, we aim to find the most frequent terms that are used repeatedly to reveal consumers' concerns for the product and their experience regarding the online purchase at Amazon.com. Therefore, we extract the most frequent terms in both helpful and non-helpful reviews by considering the results from the BOW technique, bi-grams model, and tri-grams model. We introduce the uni-gram model as BOW which assumes each word w_i is produced independently of the other words, and we use TF-IDF vectorizer to calculate the frequency of each word within the document. Furthermore, we expand the size of window of width n words over text, where n=2 is referred to bi-grams and n=3 is referred to tri-grams. The benefit of using

n-grams models is the sequence of terms can be compared to each other in an effective manner. We use the results of this technique as features to enhance the classifier performance. Figure 3.15 shows the results of this comparison for both helpful and non-helpful reviews based on their polarities.

Ranking	Non-Helpful Reviews' keywords					
	Bag of word		Bi-grams (Non-Helpful)		Tri-grams(Non-Helpful)	
	Positive	Negative	Positive	Negative	Positive	Negative
1	Like	Bad	Gorgeous, Satisfied	Wrong, Awkward	Flattering, Impressed, Cute	Smell, Worn, Ached
2	Love	Horrible	Stunning, Classic	Bad, Waste	Beautiful, Perfect, Love	Dirt, Discomfort, Terrible
3	Great	Disappointment	Amazing, Strong	Terrible, Cheaply	Nice, Easy, Return	Poor, Hated, Lack
4	Well	Smell	Elegant, Cute	Broke, Return	Loves, Recommended, Great	Irritated, Loose, Issues
5	Perfect	Junk	Shinny, Perfect	Awful, Problem	Nice, Perfect, Fashionable	Ridiculous, Hassle, Awful
6	Good	Terrible	Super, Cute	Allergic, Disappointed	Fantastic, Enjoy, Satisfied	Disappointed, Smelling, Terribly
7	Nice	Disgusting	Perfectly, Great	Stuck, Return	Solid, Sharp, Strong	Difficult, Waste, Ripped
8	Top	Sad	Sturdy, Worth	Rip, Hate	Comfy, Well, Gorgeous	Wrinkles, Loss, Bulky
9	Best	Uncomfortable	Soft, Adorable	Poor, Destroy	Adorable, Roomy, Like	Awkward, Bent, Upsetting
10	Happy	Broke	Worked, Well	Weird, Broke	Cool, Inexpensive, Best	Problem, Dirty, Broken

Figure 3.15: Top 10 high-frequent terms in different models for non-helpful reviews

The highlighted terms for the BOW technique are very general compared to those in the bi-grams and tri-grams models. For instance, "Nice" is a very general adjective to describe a product, but when it is accompanied by "perfect" and "fashionable", the phrase for describing the product gains more value. In addition, the terms in non-helpful reviews are still too general to be reliable for future purchases, which explains why they do not attract consumers' votes.

Moreover, we extract the top ten high frequent positive and negative terms for the helpful reviews by considering the results from all three techniques that we discussed earlier for non-helpful reviews. Figure 3.16 shows the results of the top ten high frequent terms for helpful reviews.

Ranking	Helpful Reviews' keywords					
	Bag of word		Bi-gram (Helpful)		Tri-gram(Helpful)	
	Positive	Negative	Positive	Negative	Positive	Negative
1	Like	Bad	Stylish, Soft	Bent, Return	Comfortable, Perfect, Stability	Cheap, Disappointed, Worse
2	Love	Horrible	Light, Handy	Junk, Stiff	Recommend, Sufficient, Appropriate	Falling, Suffer, Return
3	Great	Disappointment	Fine, Comfortable	Worn , Wrinkles	Durable, Supportive, Flexible	Wrong, Worn, Pain
4	Well	Smell	Gorgeous, Classic	Cheap , Difficult	Love, Amazing, Perfect	Uncomfortable, Sweaty, Ripped
5	Perfect	Junk	Pleased , Reasonable	Uncomfortable, Disappointed	Accurate , enough, like	Waste, Worn, Difficulty
6	Good	Terrible	Flattering, Pretty	Waste , Allergic	Refunded, Wonderful, Pleased	Fake, Horrible, Smell
7	Nice	Disgusting	Recommend , Easy	Rip , Sting	Classy, Prefer, Excited	Worn, Itchy, Loose
8	Top	Sad	Sturdy , Impressed	Wrong , Loose	Comfy, Fine, Precise	Issues, Odd, Return
9	Best	Uncomfortable	Soft , Awesome	Smell, Concern	Adorable, Stunned, Love	Pricey, Disappointment, Ugly
10	Happy	Broke	Adorable, Warm	Overpriced, Trouble	Stylish, Convenient, Return	Trash, Bad, Terrible

Figure 3.16: Top 10 high-frequent terms in different models for helpful reviews

Figure 3.16 indicates that when using the BOW technique, “Like”, “Love” and “Great” are repeated frequently in positive reviews, and “bad”, “horrible” are repeated in negative reviews, while more specific words related to Amazon Fashion such as “Stylish, Soft”, “Light, Handy”, and “Comfortable, Fine” from positive reviews are repeated in helpful reviews in bi-grams model. By contrast, negative words emphasizing the quality of products, such as “bent, return” and “junk, stiff” have a higher rank in the bi-grams model. In the tri-grams model, words with the highest rank for positive and negative reviews are more specifically related to the fabric used in products. Unexpectedly, the approach related to price does not rank first among all positive and negative reviews rated as helpful. Instead, the applicability of the product seems to be the priority for most customers, since the words “Comfortable” or “Durable” are at the top of the list.

We can also observe that in the fashion industry, customers value aesthetics, since combinations of words like “Classic”, “Stylish” have a higher frequency in the bi-grams model. Besides, product quality seems to appeal to many customers. In general, moving from the left side of the table to the right side of it shows how terms are changing from very general expression to more specific and useful explanations.

From the top 10 negative words list, we also observe that if a fashion product looks cheap, consumers consider that as a signal for a negative review. We can also conclude that “return” is a strongly negative word that makes negative reviews helpful for consumers when they are going to spend more money on purchasing products that have high valuation uncertainty. To uncover the most frequent terms

that appear in helpful reviews, we extract the features of using bi-grams for the positive and negative words to create a network of terms. Figure 3.17 shows the network of positive words by applying bi-grams model. We also have the network of terms for negative terms which is shown in Figure 3.18.

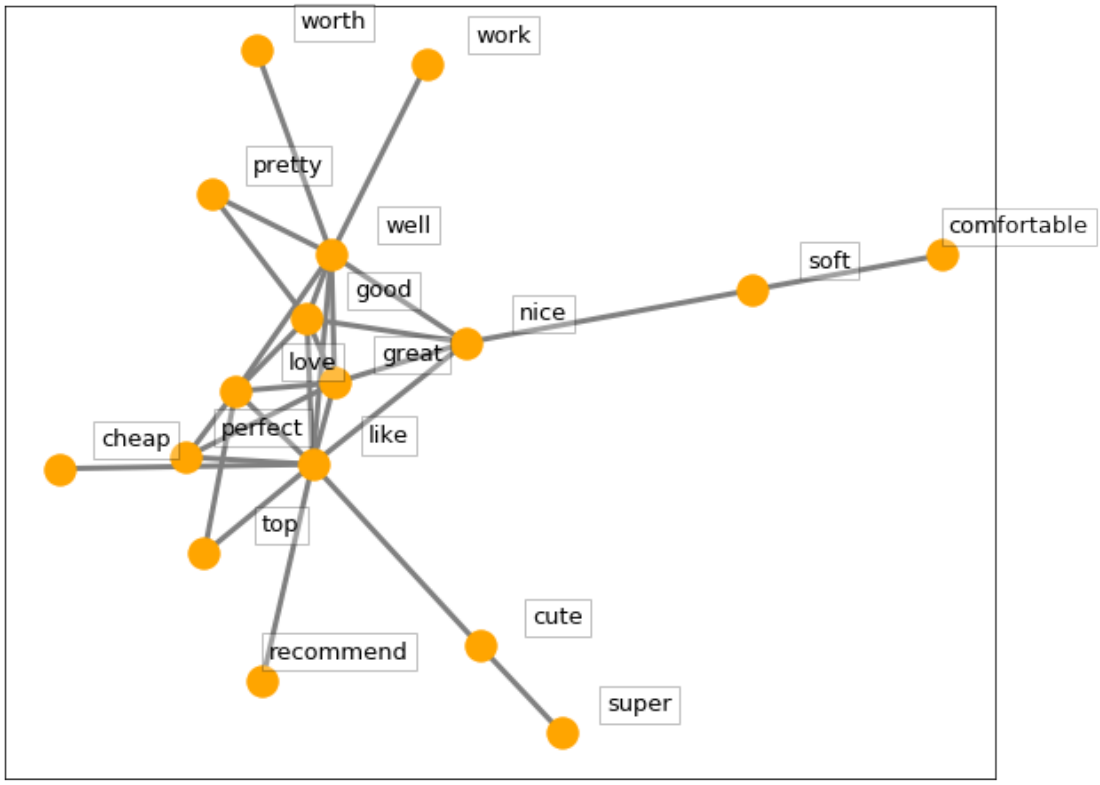

Figure 3.17: Co-occurrence network of terms for positive terms

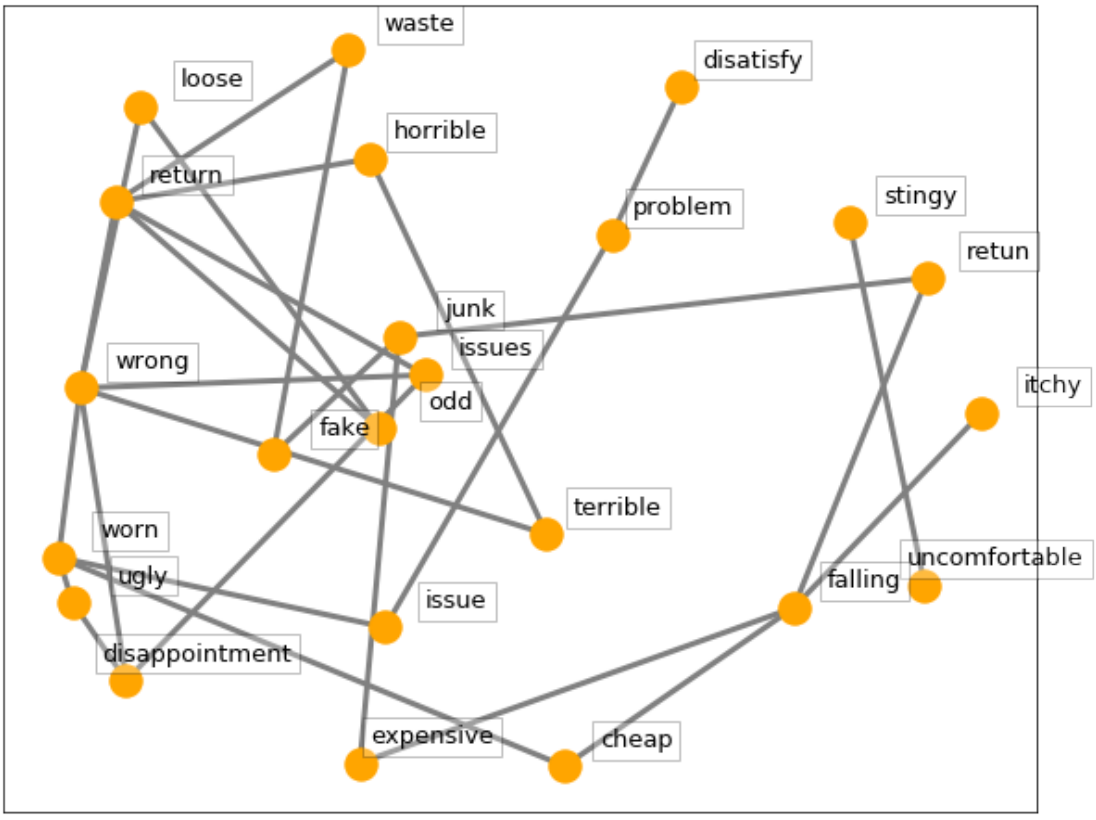

Figure 3.18: Co-occurrence terms of network for negative terms

3.6 Support Vector Machine (SVM)

We begin this section with a formal definition of support vector machine (SVM). Then, we discuss the results of examining two categories of features including: (i) topics analysis (by using LDA model) and (ii) topics plus bi-grams with TF-IDF vectorizer for enhancing the accuracy of prediction. We evaluate the results by comparing the difference between the impact of using each category of features on the algorithm's performance. We use an SVM for the text classification, which is one of the most powerful supervised algorithms in this field. SVM was originally developed as a binary linear classifier similar to logistic regression. We use SVM in this study because sentiment analysis is a binary classification, and it can work with huge datasets. In this research, to train the classifier, we employ a

manually created training set. SVM input is the score of the opinion word about a feature we defined in a review. We aim to extract the features at reviews and evaluate reviews' helpfulness to give the score to them. We use the scores to classify them as helpful and non-helpful reviews.

3.6.1 Topics Analysis (LDA) for SVM classifier

We examine the topics analysis by using LDA to select features to find the semantic structures. Then, we use the SVM classifier based on the topics-text matrix, which is shown in Figure 3.13. Text representation capabilities and feature reduction of the LDA model can improve the performance of SVM classification. In other words, each document is generated by 15 topics from topic zero to topic fifteen, and the frequency of each word is observed and is filtered in each document. We train the SVM classifier by using topics that are generated by LDA as the inputs.

3.6.2 Topics + Bi-grams TF-IDF vectorizer for SVM classifier

We add bi-grams by using TF-IDF algorithm to the topics and use the outputs as features for SVM classifier. We first use the bi-grams with TF-IDF technique to extract terms from the documents by weighing the appropriate term, and we use the LDA's outputs as topics to model the text probability. Therefore, each topic demonstrates the probability distribution of words in the document. Using the bi-grams TF-IDF technique reveals the importance of the words in the document, thereby distinguishing it from other documents and unveiling the most improtant

and relevant terms in each topic. We represent the performance of the topics analysis (LDA) and the topics plus bi-grams TF-IDF vectorizer model in the SVM classifier and compare the results of the evaluation in the next section.

3.6.3 Confusion Matrix

In text classification, we evaluate the correctness of the classifier predictability. The statistics that we are looking for to compare the results in different approaches are precision and recall in addition to accuracy and F-measure. The statistics are shown on the confusion matrix, which present all the relevant information for each approach. The confusion matrix is presented as follows:

Table 3.2: Confusion Matrix

Actual class	**Predicted as helpful**	**Predicted as non-helpful**
Helpful	True Helpful (TH)	False Non-Helpful (FN)
Non-Helpful	False Helpful (FH)	True Non-Helpful(TN)

The parameters of this matrix are described as follows:

1. **True Helpful** : sample is belonging to the helpful class predicted as helpful
2. **True Non-Helpful**: sample is belonging to the non-helpful class predicted as non-helpful
3. **False Helpful** : sample is belonging to the non-helpful class predicted as helpful
4. **False Non-Helpful** : sample is belonging to the helpful class predicted as non-helpful

Furthermore, we have the evaluation metrics that are mentioned above, and we computed them based on the values in the confusion matrix. The first metric is accuracy which shows the number of correctly predicted helpful reviews out of all the reviews, as shown in the following equation:

$$Accuracy = \frac{TH + TN}{TH + TN + FH + FN} \tag{3.3}$$

The second metric is precision, which is the number of true helpful reviews out of all the reviews that are either predicted correctly as helpful reviews or assigned as helpful reviews. The following equation shows how we calculated it:

$$Precision = \frac{TH}{TH + FH} \tag{3.4}$$

The third metric named as recall, which is the number of true helpful reviews out of the actual helpful reviews, and it is given by

$$Recall = \frac{TH}{TH + FN} \tag{3.5}$$

Finally, the last metric is F-measure, which is a weighted method of precision and recall metrics, and it is calculated as follows

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{3.6}$$

The value of F-measure is between zero to one and shows better results when

it is closer to one. Moreover, to present the performance of our proposed text classification algorithm, we compare the result of applying two categories of features on SVM classifier.

3.7 Results

The result for this study has two subsections including performance analysis and cross-validation analysis. In performance analysis, we explain how each category of features generate different accuracy as well as other performance metrics. In cross-validation technique, we evaluate the models by training several models on the subset of dataset and evaluate them on the complementary subset of the data. We split the training data into five-folds and evaluate the accuracy of each fold and then take the average of them. Furthermore, we examine the validation of our models by checking the receiver operating characteristics (ROC) to evaluate the overall performance of our models.

3.7.1 Performance Analysis

In this section, we present the results from comparing the performance of topics by using LDA for SVM and topics by using LDA plus bi-grams by using TF-IDF vectorizer for SVM. As it shown in Figure 3.19, the performance of helpfulness prediction related to the helpful reviews for the topics plus bi-grams TF-IDF vectorizer for SVM algorithm is slightly higher than the performance of only topics for SVM algorithm. Figure 3.19 also shows the comparison between the results of confusion matrix for both topic analysis and topics plus bi-grams TF-IDF

vectorizer on SVM algorithm for all four metrics.

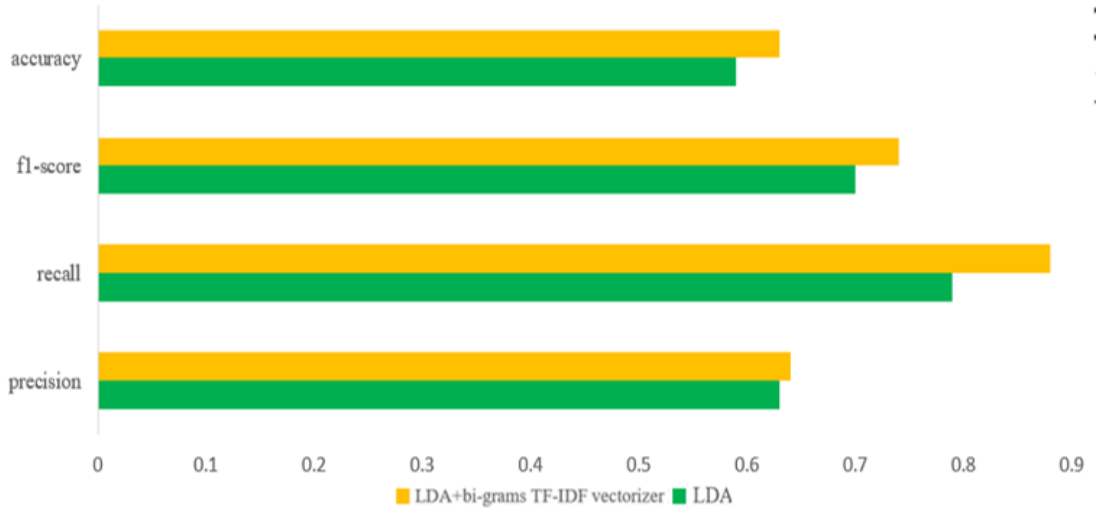

Figure 3.19: Performance analysis based on confusion matrix

As can be inferred from Figure 3.19, the topics plus bi-grams TF-IDF vectorizer on SVM algorithm has higher accuracy compared to only topics on SVM algorithm. Both these approaches emphasized on the helpfulness of reviews, review rating, and review content features. In the following section, we use the k-folds cross-validation analysis to evaluate the validation of the models.

3.7.2 K-Folds Cross-Validation (KCV) on SVM classifier

In this study, we applied k-folds cross-validation (KVC) technique to the SVM classifier and compared the results for five-folds. The literature shows that k-folds cross-validation (KVC) procedure is simple, effective, and reliable (e.g., Anthony and Holden (1998); Liu and Liao (2017); Zhang and Wang (2016)). In this research, we split the dataset in five independent subsets, and one of the subsets is used to train the SVM classifier. The Figure 3.20 shows how the five-folds for the training dataset works. We evaluate the results for both categories of features on

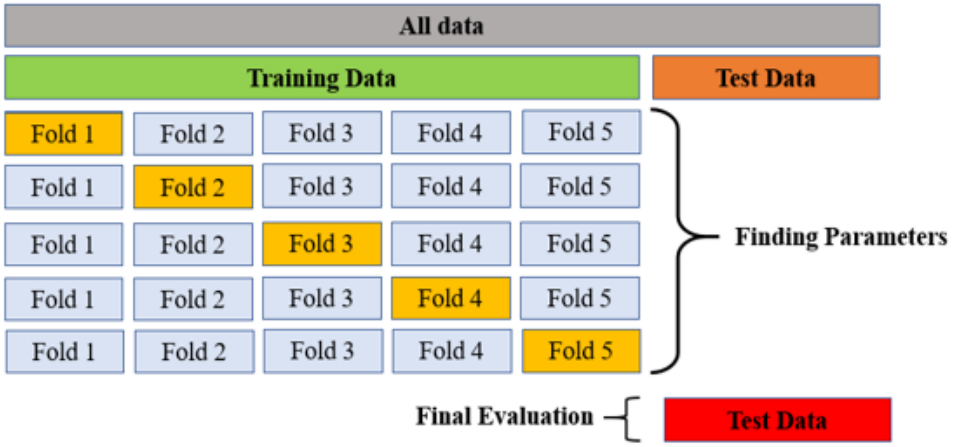

Figure 3.20: K-folds Cross-Validation technique with K=5

SVM classifier by applying the KVC which is shown in Figure 3.21. The results of

Cross-validation Accuracy for SVM						
	Fold #1	**Fold #2**	**Fold #3**	**Fold #4**	**Fold #5**	**Average**
Topics Analysis (LDA)	58.29%	57.64%	57.46%	53.78%	54.98%	56.43%
Topics plus Bi-grams TF-IDF	75.61%	76.47%	68.77%	71.17%	64.52%	71.31%

Figure 3.21: KVC technique on SVM classifier

applying KVC on SVM classifier shows that topics plus bi-grams TF-IDf vectorizer features increases the accuracy of SVM classifier compared to only using topics analysis on SVM classifier. We also check the performance of each model by using ROC model on KVC. Furthermore, we evaluate the overall performance of these two classification models by the area under the Receiver Operating Characteristics (ROC) curve by measuring the true helpful rate on the y-axis and false helpful rate on the x-axis. The equations for the true helpful fraction and false-helpful

fraction are as follows:

$$\text{true-helpful fraction} = \frac{TH}{TH + FN}$$

and

$$\text{false-helpful fraction} = \frac{FH}{FH + TN}$$

The functionality of the ROC curves for the topics plus bi-grams TF-IDF vectorizer on SVM model and topics analysis on SVM are shown in Figure 3.22. The ROC curve plots true-helpful fraction versus the false-helpful fraction at different classification thresholds. In other words, the greater area under the curve means the better performance of the classifier. As it is shown in Figure 3.22, the area under the topics plus bi-grams TF-IDF vectorizer on SVM model is larger than the area under the topics analysis on SVM model.

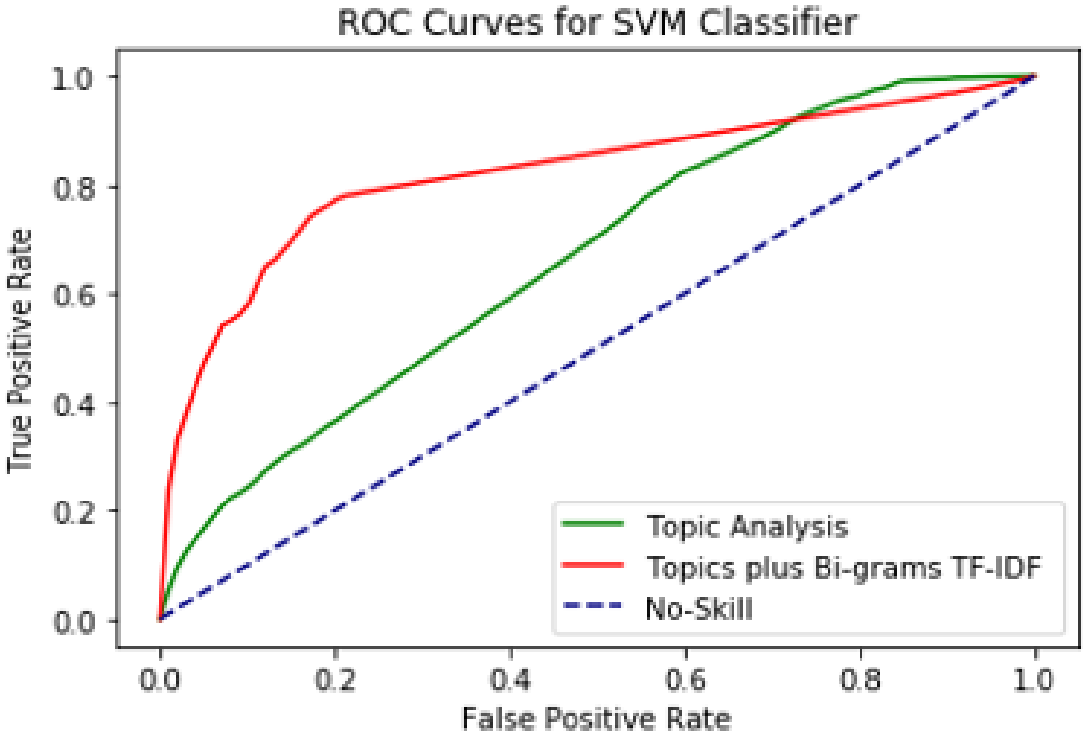

Figure 3.22: ROC with KVC technique for SVM classifier

3.8 Conclusion and Future Work

In this study, we developed an innovative method that can predict the helpfulness of reviews and identify the most frequent terms used in helpful and non-helpful reviews of fashion product purchases on an online platform. We enhanced the SVM classifiers' performance by using appropriate text preprocessing methods, including cleaning data by regretting those reviews that are written on non-English languages. We applied topics analysis by using LDA model for the reviews and obtained the best estimation of the number of topics by choosing the fifteen topics. We added another approach by using topics plus bi-grams TF-IDF vectorizer to enhance the performance of the classifier. Experiments in our study show that the accuracy of the topics plus bi-grams TF-IDF vectorizer on SVM model is higher than the accuracy of applying only topics on SVM model in predicting the

helpfulness of the reviews.

We evaluated the accuracy, precision, recall, and F1 score of sentiment classification for both categories of features on SVM algorithm. Furthermore, we examined the possible reasons of gaining more helpful votes, including longer reviews, rating system, and specific terms within a review. We found that longer reviews are not necessarily helpful ones and that the rating system does not have a strong correlation with the helpfulness of reviews.

We also find that helpful reviews are neither strongly positive nor strongly negative. The helpful reviews are balanced with the use of certain terms related to product quality and consumers' suggestions on returning or purchasing products, which make them helpful for trusting the reviews and voting for them. We developed a model that provides both consumers and online retailers with a very stable environment in which to interact without having issues about fake reviews with manipulated helpful votes. Our analysis helps to extract specific features from any set of reviews for any sensory category of product sold on the online channel.

- **Online retailers** can make good use of our model to obtain essential information for selling their products once they have received a certain number of reviews to sell their product. This analysis can identify helpful reviews without focusing on the number of votes or having a verified tag.

- **Customers** can trust the reviews irrespective of whether they have a verified tag or a large number of votes. Moreover, they can choose their favorite

a wide variety of reviews that are either unvoted or higher voted reviews.

For future work, we would like to extend this study to different fields offering sensitive service or products to consumers through online channels, and where consumers' reviews are key to the purchase decision making process. We seek to use different vectorized features and structured features for the models. Furthermore, in this study, we used our model for binary classification problems (i.e., helpful reviews vs. non-helpful reviews); but we can extend the model to address multi-classification problems in future work.

Chapter 4

Sentiment Analysis on Tweets across the development of COVID-19

4.1 Introduction

The coronavirus disease known as COVID-19 started in December 2019, when several patients from Wuhan Hubei province in China reported severe health symptoms. Since then, COVID-19 has spread significantly to many countries and is now considered a pandemic. According to the World Health Organization (WHO) report in July 2020, 10,533,799 cases have been confirmed. Moreover, the report showed that more than half a million deaths have resulted across the world [1]. However, the WHO solution for reducing the number of confirmed cases and mortality rates is isolation and self-quarantine, which has led to the biggest lockdown in the history.

Spending time at home and searching for the news has become one of the leading forms of entertainment as a result. Spending time at home and searching for the COVID-19 related news has become one of the leading forms of entertainment as a result. Twitter has become one of the most significant ways of sharing

information and expressing feelings regarding COVID-19 during this pandemic.

Twitter users can forward each tweet to their network, which is referred to as retweeting, and speed the information sharing process. Thus, retweets can represent Twitter users' interests on a large scale and the popularity of tweets based on their content and the volume of retweets. However, during the COVID-19 pandemic, sixty percent of the misleading information on Twitter remained on the platform as a "resource" for users to retweet [2]. The research objectives for this study are (i) predicting the popularity of tweets based on the volume of retweets, (ii) selecting the features including n-grams and topics that help to understand the public's sentiment at different stage of COVID-19, and (iii) building a model with the highest accuracy to predict the popularity of the tweets. In this research, we select the features that have a significant impact on the performance of the random forest (RF) classifier, support vector machine (SVM) classifier, and logistic regression (LR) classifier for predicting the popularity of tweets based on their content.

We analyze almost three hundred thousand tweets on COVID-19 that are written in English to evaluate each tweet's content and to categorize users' fears, anger, hope, and any expressions of racism related to the pandemic from all over the world. We then select features from (i) topics analysis by using the latent Dirichlet allocation (LDA) technique, (ii) n-grams by using TF-IDF vectorizer, and (iii) topics analysis plus bi-grams TF-IDF vectorizer. Then, we use the features for the selected classifiers to improve the prediction performance of tweet

[2]"On Twitter, almost 60 percent of false claims about coronavirus remain online" *The Washington Post* April 7,2020

popularity. We compare the result of the classifier with different categories of features to highlight the difference in results and choose the best performance of all classifiers. We find that RF has the highest accuracy among the other classifiers and topics analysis plus bi-grams TF-IDF vectorizer has the highest impact on the performance of each classifier.

We organized this study be reviewing the literature in §4.2, and specifically reviewing the background of the impact of social media and Twitter on the pandemic. In §4.3, we introduce the dataset by having overview of tweets in §4.3.2 and overview of the retweeted tweets in §4.3.2. In §4.4, we clean the tweets, and we define the binary response variable for the popularity of tweets based on the volume of retweets. In §4.5, we discuss the process of selecting the features in the dataset and explain it in detail by having two subsections including §4.5.1, and §4.5.2 that presents the n-grams with TF-IDF vectorizer. §4.6 introduces the supervised machine learning algorithms in this study by having three subsections §4.6.1, §4.6.2, and §4.6.3. We have the model validation in §4.6.4, and the results in §4.7. Finally, we have the conclusion in §4.8.

4.2 Literature Review

Our research contributes to two research streams, including the impact of media, and particularly Twitter during pandemics, and retweeting behavior based on the tweets' content. With regard to the first research stream, Odlum and Yoon (2015) studied the use of Twitter during the Ebola outbreak to monitor information sharing among users and examine the users' behavior and their knowledge

of the disease during the pandemic. The result of this study revealed the pattern of spread of the information among the public and highlighted the value of Twitter as a tool for supporting public awareness. Lazard et al. (2015) studied a textual analysis to examine the public's concerns about the Ebola virus and safety information. The study highlighted the efficiency of Twitter in public heath communication. Jain and Kumar (2015) studied the use of Twitter in the 2015 H1N1 pandemic (known as Swine flu) to make an inspection system by analyzing the relevant information related to Influenza (H1N1) and enhancing the public awareness of it in India. They studied public opinion regarding H1N1 flu and analyzed the tweets, and classified them as relevant and irrelevant. Their results highlighted the importance of social media for tracking a disease.

Szomszor et al. (2011) analyzed the tweets and online media related to the Swine flu pandemic of 2009 with the aim of identifying the popularity of true information. They found that poor scientific knowledge can still be shared in public and cause harm. Furthermore, there are several studies that have examined the Twitter content during pandemics in order to analyze how the public express their feeling in the early stages of the disease (e.g., Ji et al. (2013, 2015); Mamidi et al. (2019)). The second research stream is related to retweeting behavior. There are several studies that have contributed to this filed by offering a solution for predicting the results of important events such as games, and political elections with the support of retweet volume (e.g., Hong et al. (2011); Suh et al. (2010); Yang et al. (2010)).

Some of the research into this aspect of retweeting has examined the reasons

why users retweet certain information without making an effort to predict the retweet. Boyd et al. (2010) empirically examined several case studies on Twitter to analyze the retweeting behavior, and understand why and what the users retweet. Their study highlighted that the bias in interpreting of tweet caused the spread of wrong information on Twitter. Kwak et al. (2010) studied the impact of retweeting on information sharing by ranking the users based on their number of followers and followings and compared it to the volume of retweets to evaluate the popularity of tweets. The result of this study showed the volume of retweet based on its content has a stronger impact than the number of people who follow the Twitter account's user.

Macskassy and Michelson (2011) explained that user's retweeting behavior has several factors to share a particular piece of information. They built a model to focus on users' topics of interest to understand the retweeting behavior at the individual level. Zhang et al. (2015) proposed a model that supports textual analysis and the social network for predicting retweeting behavior. They compared the performance of their model with other supervised machine learning algorithms. Naveed et al. (2011) examined the impact of the tweet's content on its retweet volume. They examined two different levels of content-based features in tweets and predicted the retweetability of the tweet. Zhao et al. (2011) developed a model by using the LDA model for short tweets and the effectiveness of the proposed model on their analysis to compare the difference between Twitter and traditional media.

We contribute to this research stream by applying topics (LDA) on classifiers as

well as topics plus bi-grams TF-IDF vectorizer models on our classifiers to enhance their performance. However, tweets with at most 280 characters bring serious challenges to the effectiveness of applying the LDA model. Mehrotra et al. (2013) empirically established a new method by using hashtags and improved LDA topic models without changing the machinery of LDA. There are also several studies related to using topic modeling and LDA over short text in different social media by either generating of word co-occurrence patterns (i.e., bi-grams, tri-grams, and uni-gram plus bi-grams) in a document or modeling the documents as a mixture of topics(e.g., Cheng et al. (2014); Li et al. (2016); Yan et al. (2013)).

4.3 Dataset

To implement our study, we examined a subset of a dataset of tweets related to COVID-19. The data has over 4 million tweets in four languages, including Spanish, English, French, and Russian, from March 27th to June 5th, 2020. In this study, we focused on tweets that are written in English, which reduced the dataset to two hundred and fifty thousand tweets. Several tweets are missing information, which narrows down our dataset to almost two hundred thousand tweets. The following table shows the relevant information about the dataset and an example of one unique record. The data is imported into Python console by using numpy, nltk, and pandas packages.

In the following section, we have an overview of tweets and the definitions of each attribute related to tweets. We explain the used attributes in our research.

User ID	1245698700736
Text	RT @ CIDRAP: Virologits weigh in on novel #coronavirus in China's outbreak
Language	En
User Location	Comunidad de Madrid, Espana
Hashtags	#Coronavirus
User Statues Count	805
Retweet Count	45

4.3.1 Overview of Tweets

In this section, we discuss the definition of some attributes that are used in our research.

- **User ID**- The integer number that represents a unique identifier for the tweet.
- **Text**- A post that contains the user's opinion on Twitter's platform on certain event. The tweet can be viewed by the user's followers as well as other users who searched certain keywords.
- **Language**- The language of the tweet is identified in four different categories, and "En" refers to English in our dataset.
- **User Location**- The tweet is associated with a location on the world map.
- **Hashtags**- Certain keywords by the (#) sign make the process of searching for information easy. Twitter refers to such keywords as hashtags.
- **User Statues Count**- The number of tweets that are issued by the user id.
- **Retweet Count**- Twitter users can share the information of a tweet with

Furthermore, some of the examples related to tweets on COVID-19 are shown below:

- *Why #CCP keeps on saying the unknown cause of pneumonia? The cause is obviously related to coronavirus. Let's just call it #WARS.#CCP*
- *A novel #coronavirus is a new strain of the virus that has not been previously identified in humans*
- *I always feel weird hoping for another coronavirus outbreak to rationalize our research!*

The examples are chosen randomly without mentioning the users' names. The content is showing users' knowledge about COVID-19, their fears, and their hope about this pandemic along with hashtags. In this study, we extract the features from texts, and hashtags and retweets' content.

4.3.2 Overview of the retweeted tweets

We use the median of the dataset, which is the midpoint value for the observations to categorize the volume of retweets in this study. Thus, if the volume of retweets for a tweet is 136 times retweets or more, we consider that tweet as a popular tweet otherwise is a non-popular tweet. The purpose of this categorization is to describe the process of the binary response variable for future analysis. In Figure 4.1, we have the distribution of retweets for the tweets that are written in English.

As it is shown in Figure 4.1, the number of tweets with higher than 136 times of retweeting is close to the number of tweets with less than 136 times of retweeting

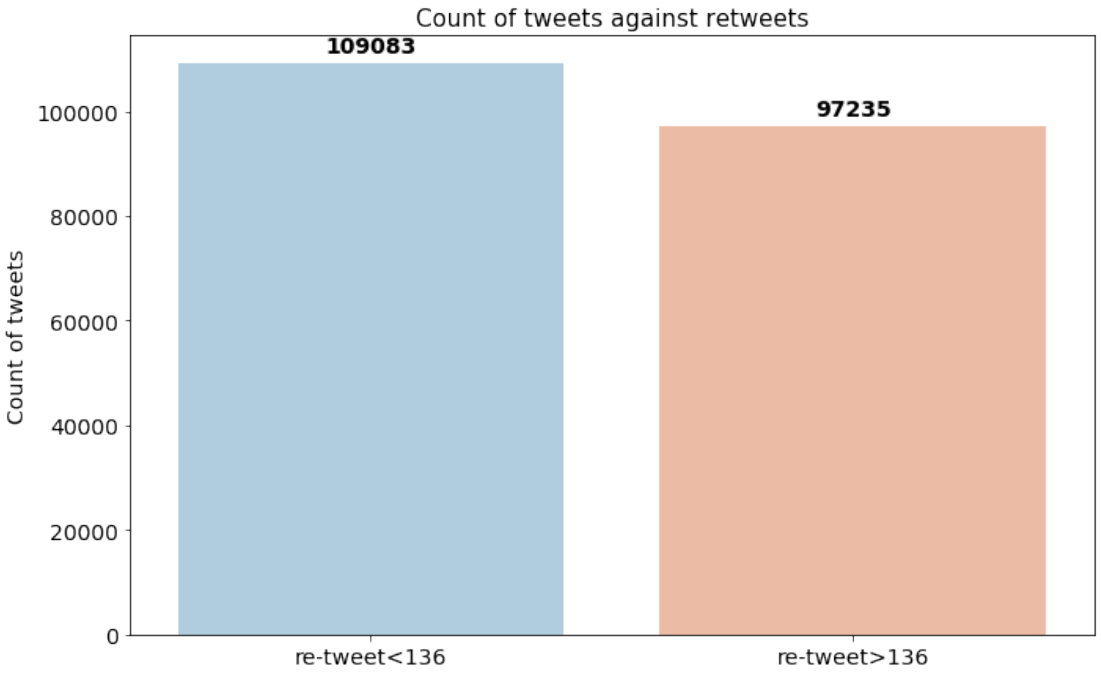

Figure 4.1: Distribution of the volume of retweets

Furthermore, the distribution of the length of tweets is shown in Figure 4.2. As it is shown, the large number of tweets have between 100 to 140 characters. Note that tweets should not have higher than 280 characters, but still, there are some tweets that poses more than the certain characters.

Furthermore, Figure 4.3 shows the frequency of tweets that have been written by the citizen of the ten countries in English. USA and Canada have the highest number of tweets related to COVID-19.

In the following section, we examine the text preprocessing on tweets, and categorize the people feeling and emotions toward the pandemic.

4.4 Tweet Preprocessing

We do the text preprocessing for the tweets by taking the following steps:

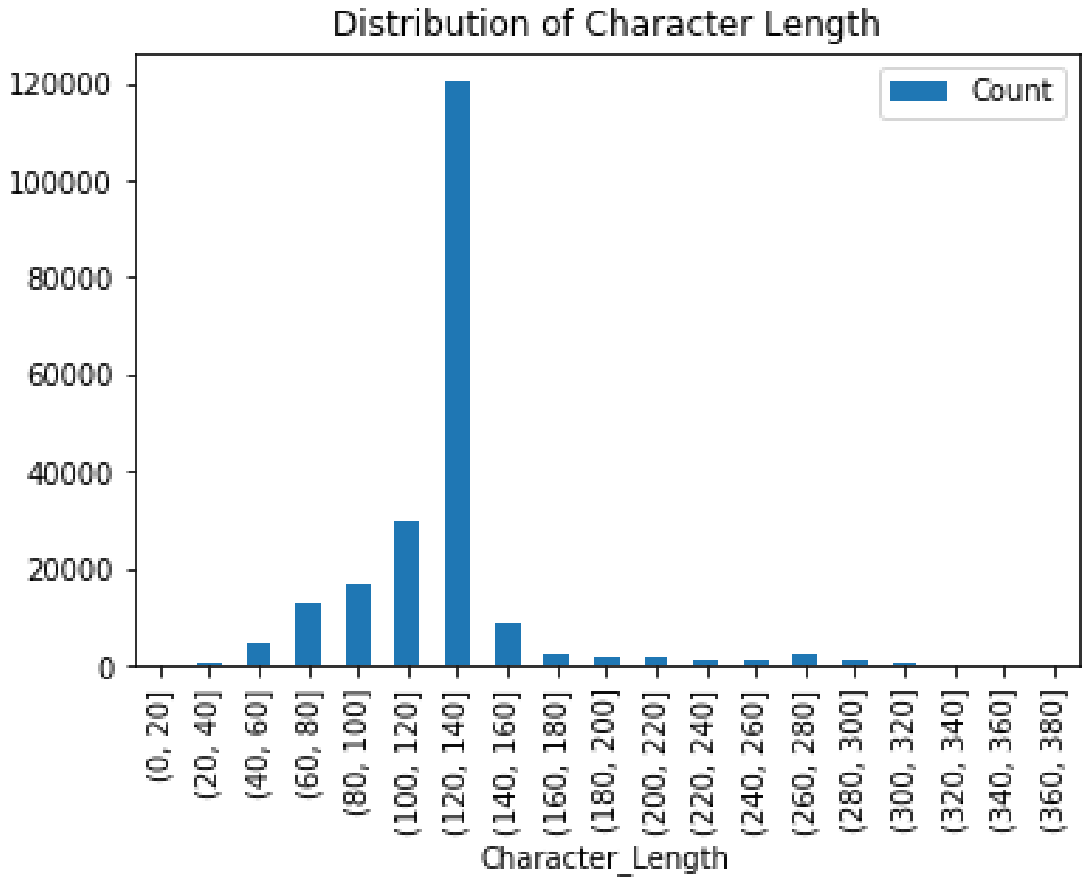

Figure 4.2: Distribution of the tweets length

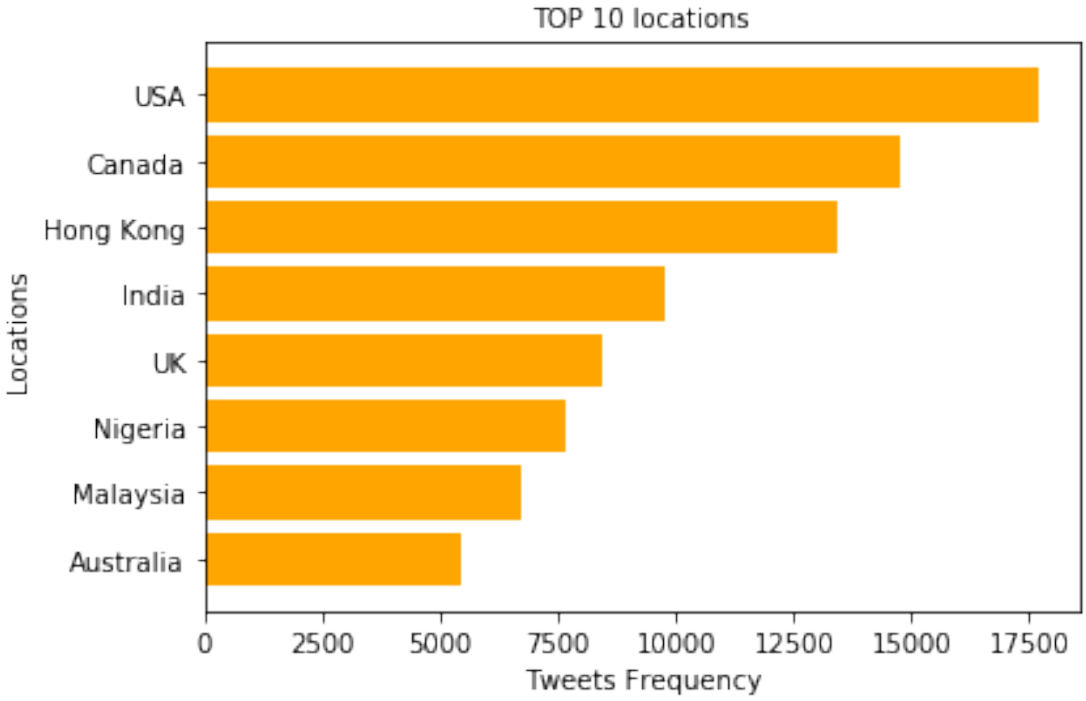

Figure 4.3: Distribution of English tweets based on locations

- convert text to the lower case.
- remove stop words (words like “the”, “of”, “in”, “at”, and punctuation).
- remove retweet keyword like “rt” and usernames like “@username”.
- remove links URL, pictures links, and emojis.
- lemmatization of the text to improve the accuracy of the analysis.

We then examine the polarity of tweets and produce the word cloud for positive and negative tweets. We examine the polarity of tweets based on lexicons of sentiment related to the words. Therefore, we identify whether (i) the word is positive or negative and (ii) how strong is the degree of positivity or negativity of the word by investigating on the sentiment metrics. We calculate the positive, neutral, and negative words within the documents and produce the compound score, which is a range between -1 and 1. In python, we use the VADER package by loading a sentiment intensity analyzer to calculate the polarity scores after cleaning the text. Hutto and Gilbert (2014) described the validation of VADER and examined its accuracy of classification for tweets into positive, neutral, and negative classes. Figure 4.4 shows the word cloud for positive tweets. As it is shown in Figure 4.4, the keywords like effect, protect, and strong are highlighted in the word cloud.

Furthermore, Figure 4.5 shows the word cloud for the negative tweets. As it is shown, most of the negative words are related to fear and anxious of outbreak and the infection.

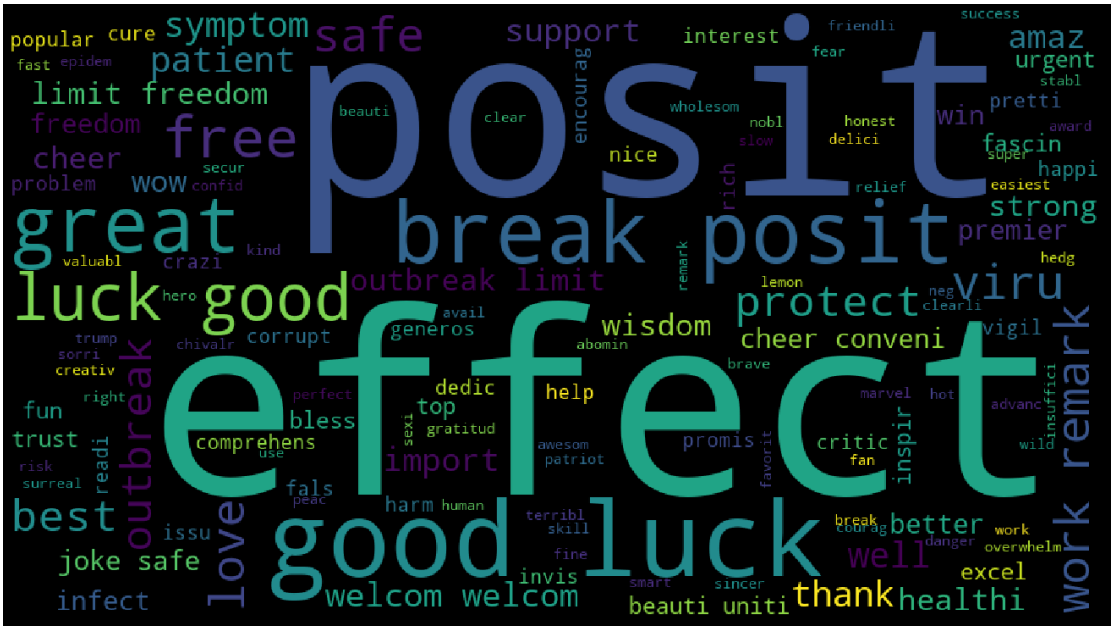

Figure 4.4: wordcloud for positive tweets

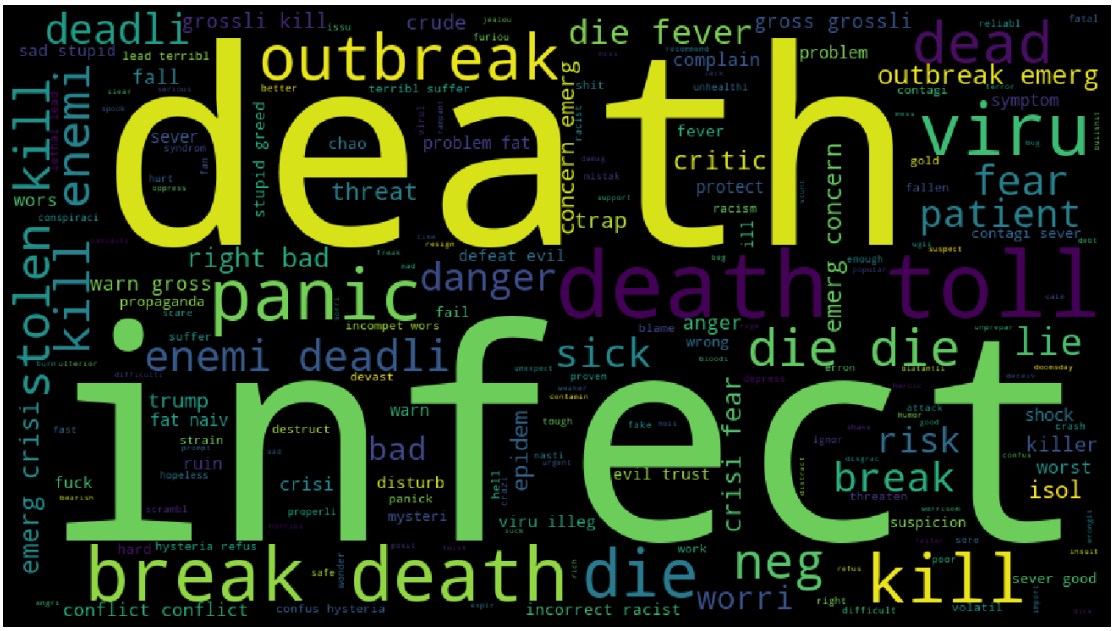

Figure 4.5: wordcloud for negative tweets

We also compare the distribution of tweets' polarity and their frequency of retweets. As it is shown in Figure 4.6, the tweets with negative contents are retweeted less than the tweets with positive content.

Categorizing the tweets based on their polarity helps us to investigate the emotions that were associated with tweets' polarity. Anxiety and fear about a new

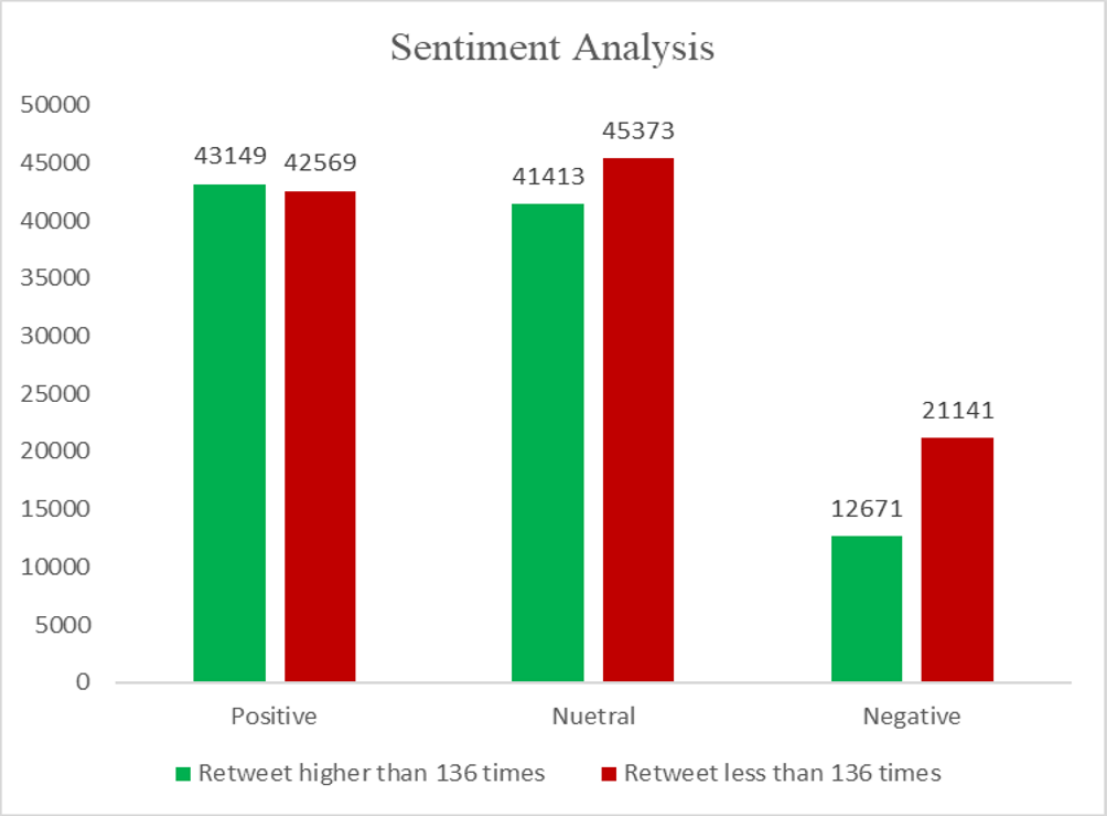

Figure 4.6: The distribution of the tweet's polarity based on their popularity

disease, having hope on the treatment and racial comments to certain countries that speed the pandemic. These emotions are the most popular emotions that are shown on users' tweets related to COVID-19. We categorize the emotions in four different clusters including; "Hope", "Fear", "Anxious", and "racism". We then import gensim package to python and create the word to vector model to learn the network of terms and detect synonymous words for all the fours clusters. Thus, in each class, the words that are associated with the emotions are extracted, which can help us for the co-occurrence network of terms for the next step.

The cluster "Hope" includes the positive sentiment about COVID-19 by having words such as heal, stable, protection, and healthy. The cluster "Fear" represents people's emotions regarding the pandemic and the primary symptom of COVID-19.

Furthermore, the tweet with anxiety is attributed to expressing the stress related to the diseases that do not have a treatment. Racism words mostly appear in negative tweets. Figure 4.7 shows the words clustering related to the aforementioned clusters in more detail. We ranked the words with the highest frequency in each cluster.

Words Clustering			
Hope	**Fear**	**Anxcious**	**Racisim**
Heal	Infected	Unexpected	Conspiracy
Stable	Fever	Havoc	Hate
Protection	Outbreak	Struck	Mania
Healthy	Worrisome	Mess	Ferociously
Improved	Vigilant	Stress	Enemies
Supproting	Dangerous	Stupidity	Terrorism
Isolation	Concerned	Risks	Gross
Strong	Contagious	Survivor	Threat
Safe	Virus	Disappinting	Trump
Clean	Warned	Revolting	Paranoid
Grateful	Scare	Suspicions	Plague
Effective	Sad	Stuck	Chaos
Heartwarming	Lack	Falling	Killer
Ready	Return	Draconian	Outrage
Defeat	Rough	Collapse	Immoral

Figure 4.7: words clustering for tweets

We also extracted the hashtags information from the tweets and ranked the top ten popular hashtags which are shown in Figure 4.8. # coronavirus has the highest frequency in tweets, and # Wuhan is in the second place since the disease started from there.

In the following section, we examine the topic modeling and latent Dirichlet allocation (LDA), and n-grams by using TF-IDF techniques to extract the features

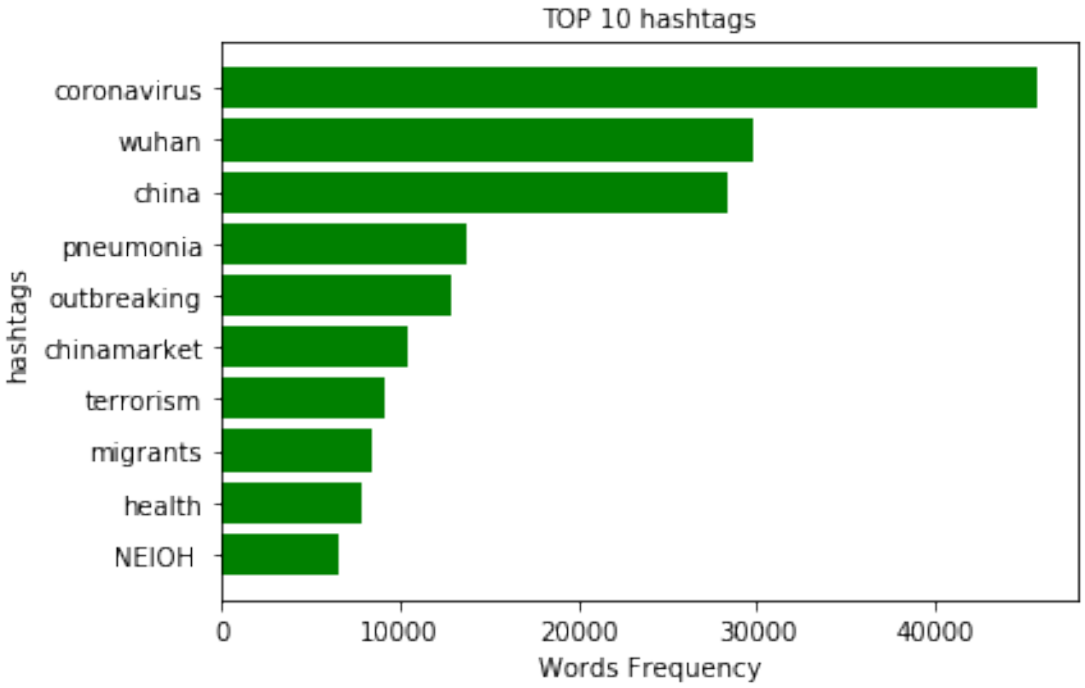

Figure 4.8: Top ten popular hashtags

from the tweets.

4.5 Features Selection

In the following subsections, we examine the different features that are used in the classifiers in order to enhance their performance. The features are including : (i) Topics analysis by using LDA model, (ii) N-grams by using TF-IDF vectorizer, and we have bi-grams, tri-grams, and uni-gram plus bi-grams, (iii) Topic analysis by using LDA plus bi-grams TF-IDF vectorizer. Our classifiers for this study are supervised machine learning algorithms including support vector machine (SVM), random forest (RF), and logistic regression (LR).

4.5.1 Topic Modeling for Short Texts

We examine the topic latent Dirichlet allocation (LDA) technique to evaluate the number of topics in the dataset. For choosing the optimal number of topics, we choose the highest coherence value score, which is shown in Figure 4.9, and it is equal to ten topics with the coherence value of 0.35. The topics generated the terms, and three first topics, along with the top ten words in each topic, are shown as follows.

- *Top 10 words for topic 0: 'fell', 'accurate', 'trust', 'tumble', 'gross', 'grossly', 'fucking', 'racist', 'cold', 'protection', 'worrisome', 'hell', 'issue', 'steal', 'healthy', 'mysterious', 'timely', 'love', 'severe', 'dangerous', 'free', 'terrorism', 'cheap', 'great', 'concerned', 'contagious', 'virus', 'warned', 'correct', 'symptoms'*

- *Top 10 words for topic 1: 'improved', 'aver', 'guidance', 'issues', 'clear', 'struck', 'scare', 'bs', 'wonder', 'welcome', 'conflicting', 'concern', 'remarkable',*

'joke', 'fever', 'better', 'interesting', 'kills', 'cure', 'lying', 'threat', 'crisis', 'killed', 'worried', 'dead', 'right', 'infected', 'trump', 'important', 'like'

- *Top 10 words for topic 2: 'anger', 'famous', 'perfect', 'bruised', 'negative', 'lead', 'lethal', 'stole', 'issues', 'exterminate', 'wrong', 'supporting', 'excuse', 'useful', 'available', 'worst', 'rapid', 'urgent', 'faith', 'warning', 'infection', 'died', 'danger', 'isolation', 'fears', 'effectively', 'like', 'emergency', 'risk', 'outbreak'*

- *Top 10 words for topic 3: 'strong', 'unconfirmed', 'difficult', 'worries', 'luck', 'vice', 'funny', 'hot', 'premier', 'critical', 'paranoid', 'worry', 'freaking', 'ready', 'bad', 'sick', 'die', 'overwhelming', 'negative', 'adversity', 'wow', 'risk', 'thank', 'patient', 'best', 'good', 'safe', 'infected', 'worse', 'positive'*

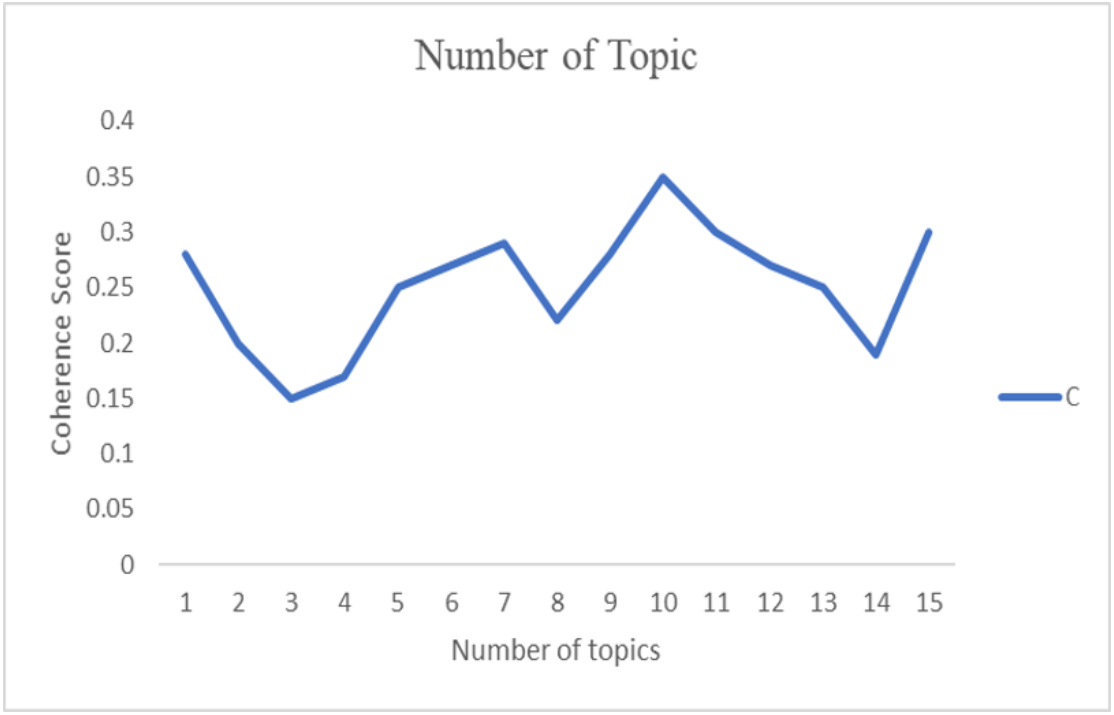

Figure 4.9: optimal number of topics

different clusters, we can create a document-word matrix. However, in order to find high-quality topics, we should have a co-occurrence network of terms. Thus, we combine the LDA analysis with co-occurrence network of terms analysis by adding the network of words "w_i" and "w_j" in the document "d". Note that w_i can have an impact on w_j and have the network. Therefore, document d is a vector of N_d words, and the matrix of topic denoted at ϕ over V vocabulary for T topics that are an extract from Dirichlet β and each topic is drawn from a symmetric Dirichlet α prior. For each cluster of words, we have z_i as the topic that generates the words and their network. Figure 4.10 shows the model.

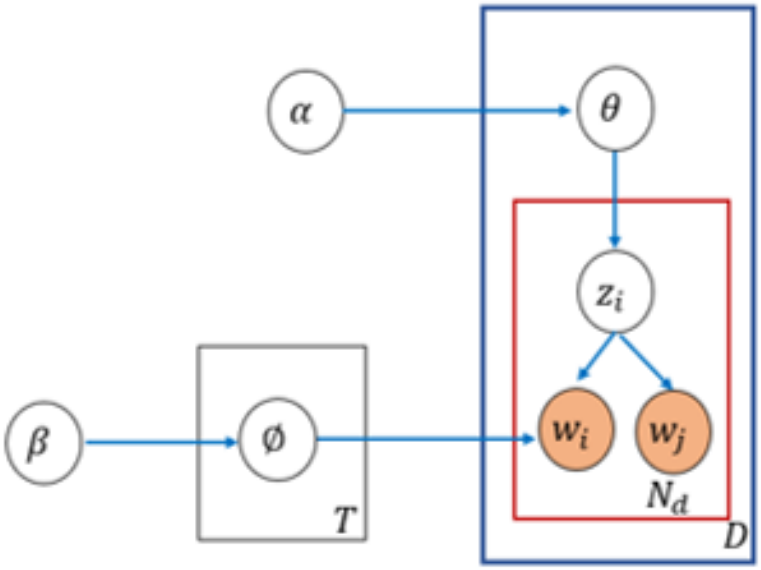

Figure 4.10: LDA model combined with co-occurrence terms of network

We have the document-word matrix in Figure 4.11 for the ten topics and the frequency of documents at each topic in Figure 4.12.

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	dominant_topic
Doc0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	**0.55**	0.05	**8**
Doc1	0.05	0.05	0.05	**0.55**	0.05	0.05	0.05	0.05	0.05	0.05	**3**
Doc2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
Doc3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
Doc4	0.03	0.03	0.03	0.03	**0.37**	0.03	0.03	0.03	**0.37**	0.03	**4**
Doc5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
Doc6	0.03	0.03	0.03	0.03	**0.52**	0.03	0.03	0.03	**0.28**	0.03	**4**
Doc7	**0.55**	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0
Doc8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
Doc9	0.05	0.05	0.05	0.05	0.05	0.05	**0.55**	0.05	0.05	0.05	**6**
Doc10	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
Doc11	0.03	0.03	0.03	**0.28**	0.03	**0.27**	0.03	0.03	**0.28**	0.03	**3**
Doc12	0.03	0.03	**0.37**	0.03	0.03	0.03	0.03	0.03	**0.37**	0.03	**2**
Doc13	0.03	0.03	0.03	**0.28**	0.03	**0.27**	0.03	0.03	**0.28**	0.03	**3**
Doc14	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0

Figure 4.11: The document-word matrix for topic analysis of tweets

Rank	Topic Number	Number of Documents
1	0	21815
2	8	3692
3	7	3558
4	3	3305
5	1	3177
6	4	3061
7	2	2430
8	9	2246
9	5	2019
10	6	1812

Figure 4.12: The frequency of documents in each topic

Each of these words has a network with the most relevant words at the highest frequency. For instance, Figure 4.14 shows that middle-aged has a network with terms like COVID-19, pandemic, coronavirus, aged 80 and over, and the United States. Another example is shown in Figure 4.15, which represents the network

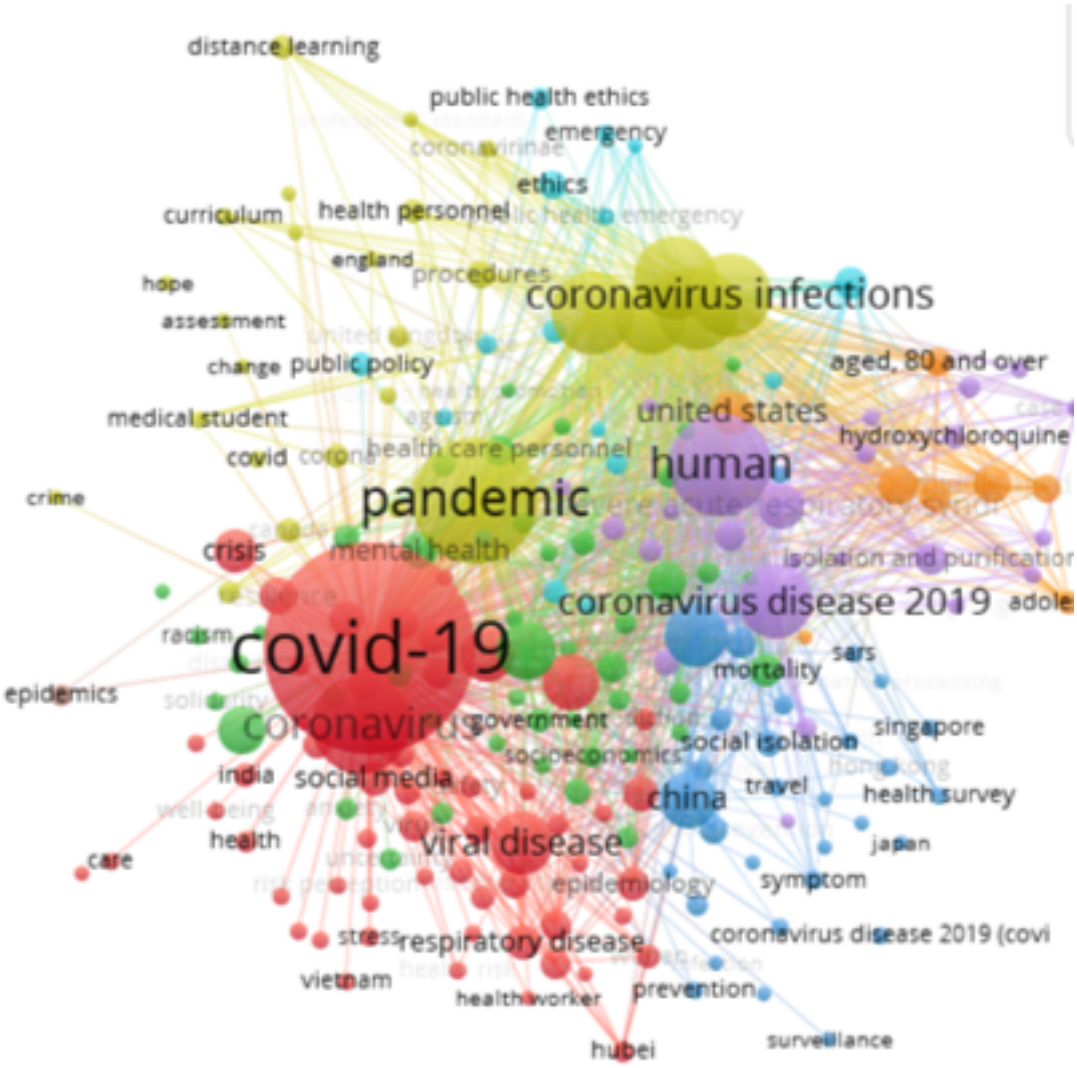

Figure 4.13: Co-occurrence terms of network for all the topics

between epidemic and COVID-19, pandemic, coronavirus disease 2019, and human. 4.13 shows the visualization of co-occurrence terms of networks for the topics as a result of topics plus bi-grams TF-IDF vectorizer model. According to co-occurrence terms of network, two words w_1 and w_2 are related to each other in topic T_i and their relation can be measured as follows:

$$R_T(w_1, w_2) = \frac{\sum_{i=1}^{n} T_i|(w_1, w_2)\frac{1}{d}}{T_i|w_2} \tag{4.1}$$

where "d" is the length of document in topic "T_i" where i shows the number of topics within the documents. According to equation (4.1), the R_T shows the relation of two words within topic T, and the strength of the co-occurrence depends on how specific w_2 is in the topic.

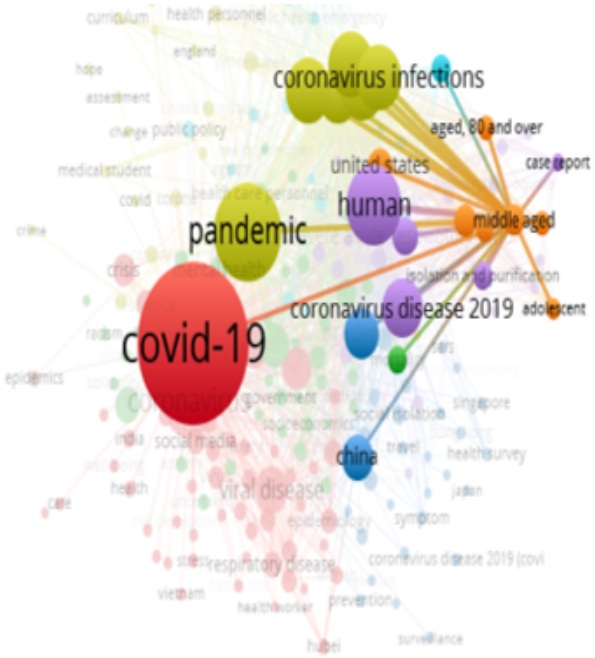

Figure 4.14: co-occurrence term of network for "middle aged"

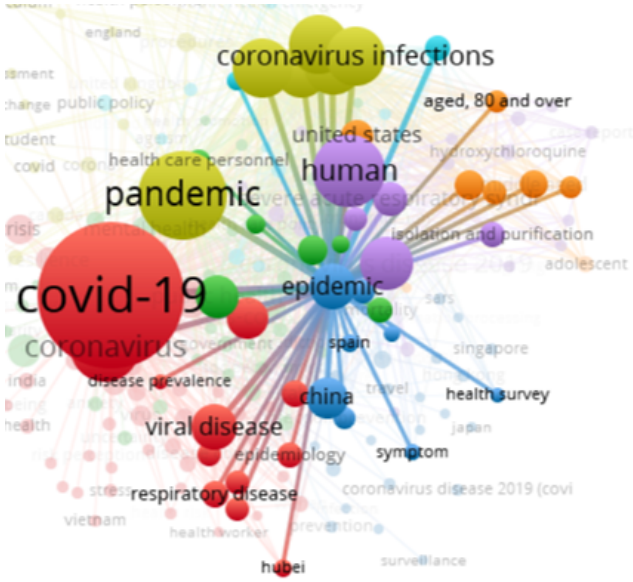

Figure 4.15: co-occurrence term of network for "epidemic"

Combining the co-occurrence network of terms with the LDA model overcome the challenge of using the LDA model for short texts like tweets. Therefore, co-occurrence network of terms generate a network of terms by considering bi-grams, tri-grams or n-grams in each topic, and TF-IDF represents the weight of the words within the documents where TF is obtained by the frequency of the term in the document and IDF supports the weight of a term to the number of documents that has that term. Therefore, we have the weight of term i in document j as follows $w_{ij} = TF_{ij} \times IDF_i$. Note that IDF_i supports the inverse document frequency of term i and TF_{ij} supports the frequency of term i in document j. Furthermore, uni-gram, bi-grams, and tri-grams are used in text mining as attributes. For instance if the sentence is “I fight COVID-19", there are three uni-grams as such “I”, “fight”, and “COVID-19”, two bi-grams “I fight”, “fight COVID-19”, and one tri-grams “I fight COVID-19”.

Creating bi-grams and tri-grams models on the dataset contributes to select features and learn what are the mos popular terms and phrases on the dataset. Tweets are short texts and using bi-grams and tri-grams can help us to find more meaningful phrases on the dataset. Figure 4.16 shows top ten popular bi-grams on the dataset and Figure 4.17 shows top ten popular tri-grams on the dataset.

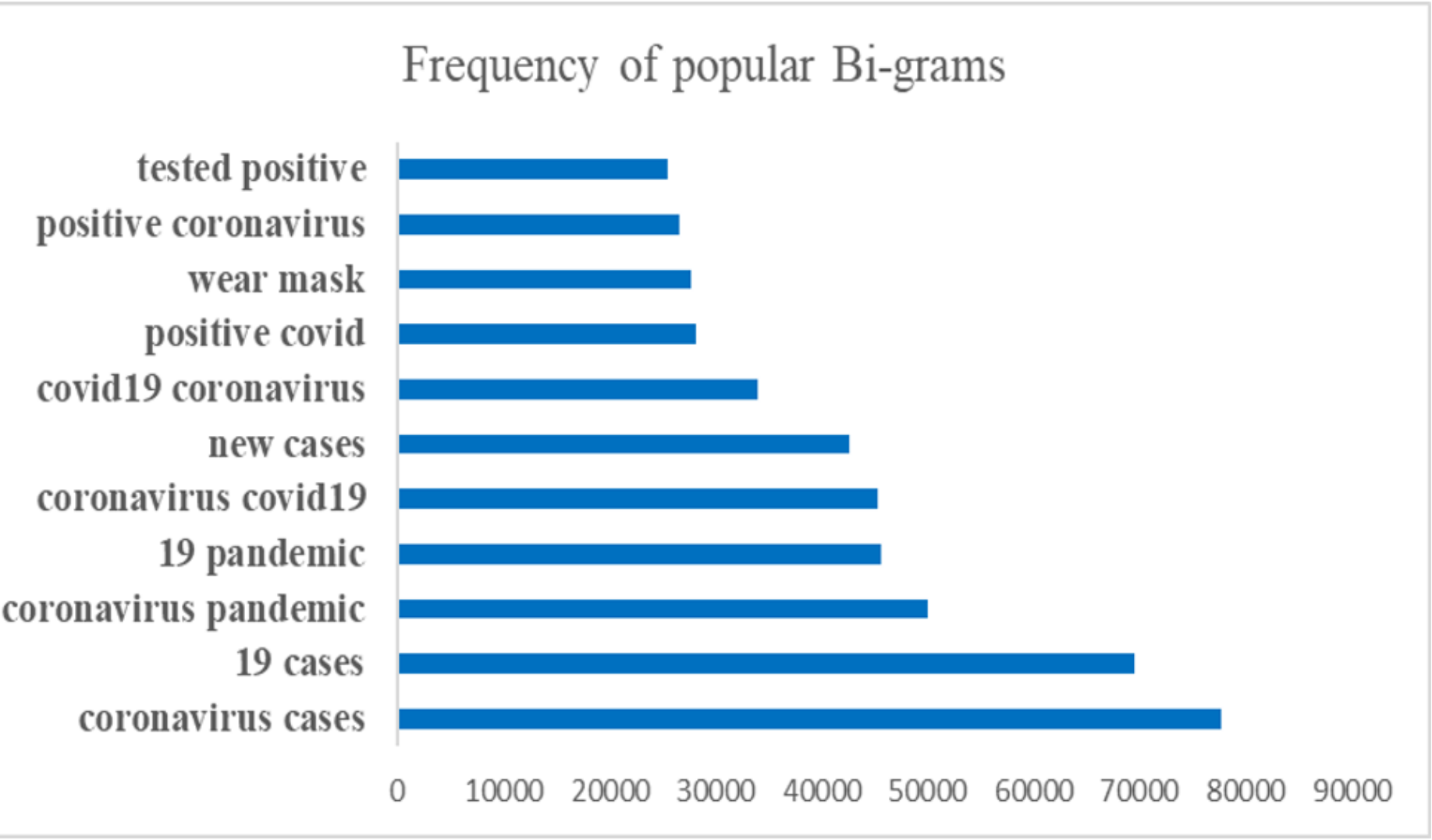

Figure 4.16: Top ten popular bi-grams for COVID-19

As shown in Figure 4.16 the bi-grams “coronavirus cases” has the highest frequency and “tested positive” is at rank ten. Moreover, “wear mask” is one of the popular bi-grams on the dataset which demonstrates the importance of fact sharing on social media. For tri-grams “COVID19 cases” is the most popular tri-grams, and all top ten popular tri-grams are related to COVID-19 or coronavirus. This reveals that people’s main concerns are related to expressing their feeling about the pandemic rather than sharing information related to how to protect themselves during the pandemic. Other popular phrases for both bi-grams and tri-grams are “000people”, “000life”, and “black life matters”. In HTML color code 000 stands for the color black, and triple zeros hashtag has become one of the popular trend during the pandemic due to the Black Live Matter events. Since the United States is one of the most active countries on Twitter, tweets on Black Lives Matter during the pandemic were also frequently repeated during the time that the dataset was collected.

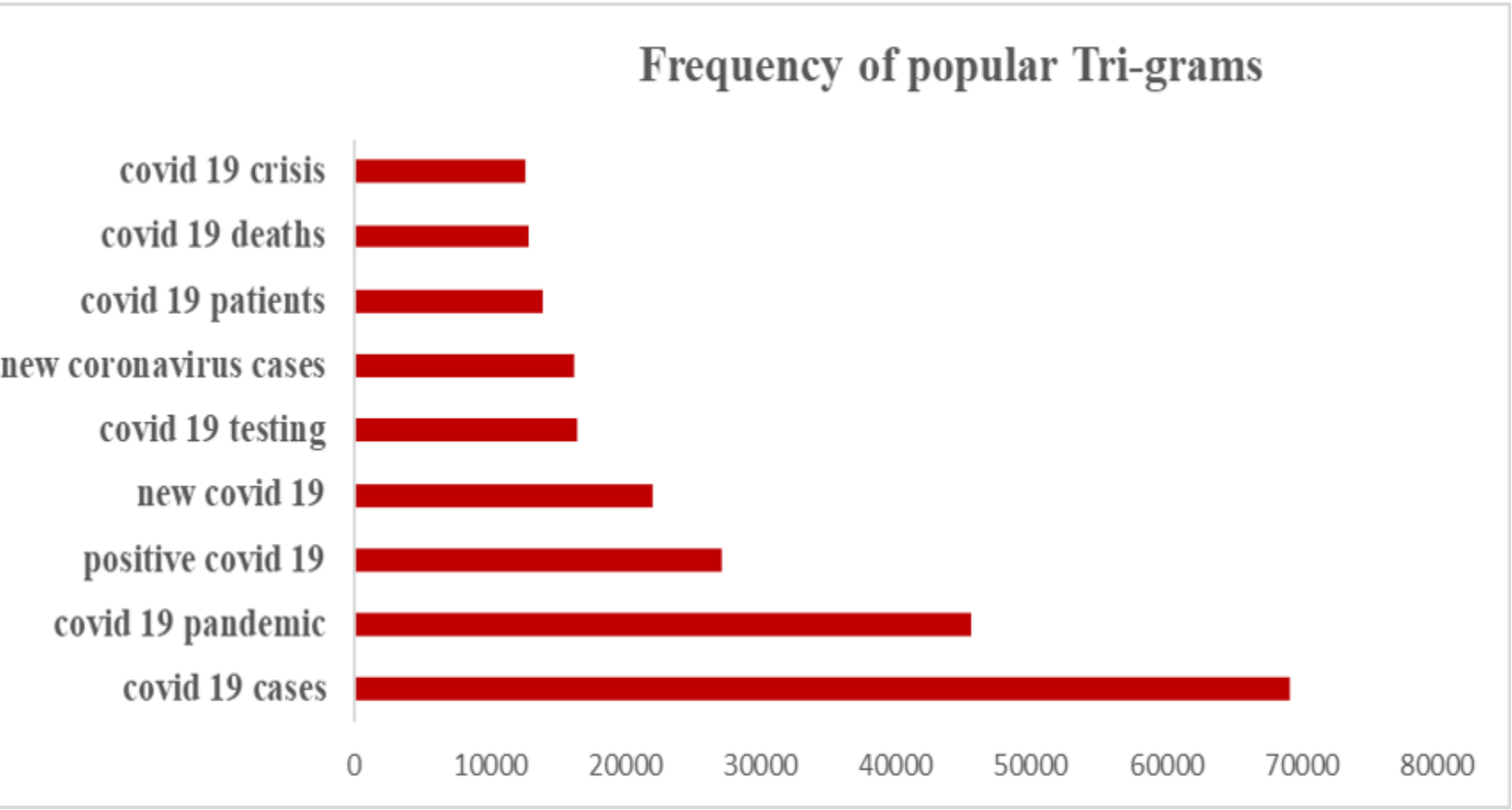

Figure 4.17: Top 10 popular tri-grams for COVID-19 tweets

4.5.2 Vectorized Features

The documents should be represented with a high level of clarity by a vector of features, and each feature should correspond to a term or a phrase in the dataset. In this study, we obtain N-grams by using TF-IDF vectorizer to meet the goal of having vectorized features. N-grams are basically a series of words or characters formed by adjusting the size of token words, as we explained in the previous section. The simplest n-grams has one word, which is called uni-gram, where n=1, and it represents the "bag of word" (BOW). Bi-grams consist of two words and n=2, representing the two-word sequence, and a three-word sequence of words is called tri-grams where n=3. Therefore, we have the following features:

- uni-gram + bi-gram TF-IDF vectorizer: a matrix with both single and paired words and their frequency and inverse document frequency within the document as a feature.
- bi-grams TF-IDF vectorizer: a matrix with paired words and their frequency and inverse document frequency within the document as a feature.
- tri-grams TF-IDF vectorizer: a matrix with three words and their frequency and inverse document frequency within the document as a feature.

4.6 Supervised Machine Learning Techniques

4.6.1 Random Forest Algorithm

Random Forest (RF) algorithm is a supervised machine learning algorithm that

random vector distributed among all trees in a forest (e.g., Breiman (1999)). The aim of this study is using (RF) for a combination of features at an individual node in order to grow a tree. The Gini Index is used in RF as a measurement for the attribute selection, and the following equation shows the index;

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{4.2}$$

where p_i is the probability that a selected object belongs to a specific class. The tree grows by using a combination of features, which is one of the main advantages of the RF classifier compared with other decision tree methods. In this study, the number of trees for RF are two hundred, and the features are topics by using LDA model, and n-grams by using TF-IDF vectorizer, and topics plus bi-grams TF-IDF vectorizer. Therefore, each case, if the dataset passes down to each of two hundred trees, and the forest choose a class with the most votes for the case. We compare the results related to LDA features as an input for RF with the combined features as an input for RF. In the following section, we have the model validation, which represents the results of both models.

4.6.2 Logistic Regression (LR)

We also considered Logistic Regression (LR) classifier for our analysis since it is the baseline supervised machine learning algorithm for classification. We use LR to classify an observation into two classes such as "popular tweets" (i.e., tweets with 136 times retweets or more) and "non-popular tweets" (i.e., tweets with less

efficient when some of the features have correlation (e.g.,Cohen and Hersh (2005); Genkin et al. (2007); Pranckevičius and Marcinkevičius (2017)). Furthermore, LR is a discriminative model and directly models the posterior probability of $P(c|t)$ by learning the input-to-output mapping by minimizing the error. LR is mainly used when the output is binary. In our study, we use LR classifier because of the need to consider two values that are related to the popularity of tweets. We consider features $\{t_1, t_2, t_3, ..., t_n\}$, and outcome c which stands for two classes and takes the value of $\{0, 1\}$ for the popular tweets and non-popular tweets. LR has a parametric form for the distribution $P(c|t_i)$, and estimates the parameters from the training data.Therefore, we have;

$$P(c = 1|t_1, t_2, t_3, ..., t_n) = \frac{1}{1 + exp(\beta_0 + \sum_{i=1}^{n} \beta_i t_i)} \tag{4.3}$$

and,

$$P(c = 0|t_1, t_2, t_3, ..., t_n) = \frac{exp(\beta_0 + \sum_{i=1}^{n} \beta_i t_i)}{1 + exp(\beta_0 + \sum_{i=1}^{n} \beta_i t_i)} \tag{4.4}$$

Notice that equation (4.4) follows from equation (4.3), since the sum of these equations is equal to one. In LR we predict the outcome to be $c = 1$, if the following condition holds,

$$P(c = 1|t_1, t_2, t_3, ..., t_n) > P(c = 0|t_1, t_2, t_3, ..., t_n)$$

substituting from equations (4.3) and (4.4), this becomes

$$1 < exp(\beta_0 + \sum_{i=1}^{n} \beta_i t_i)$$

and by taking the natural log of both sides, we have $c = 0$ if t_i satisfies

$$0 < exp(\beta_0 + \sum_{i=1}^{n} \beta_i t_i) \tag{4.5}$$

and we have $c = 1$ otherwise.

4.6.3 Support Vector Machine (SVM)

A support vector machine (SVM) is another supervised machine learning algorithm that is used for binary classification in this research. We use SVM in this study for the good reputation of this classifier on the high accuracy. The SVM finds a hyperplane to separate the positive training example from the negative one with the highest margin (e.g., Sain (1996)). The SVM classifier is memory efficient in high dimensional space which suits well in our large dataset. Figure 4.18 is the example of showing that SVM maximizes the margin around the separating hyperplane.

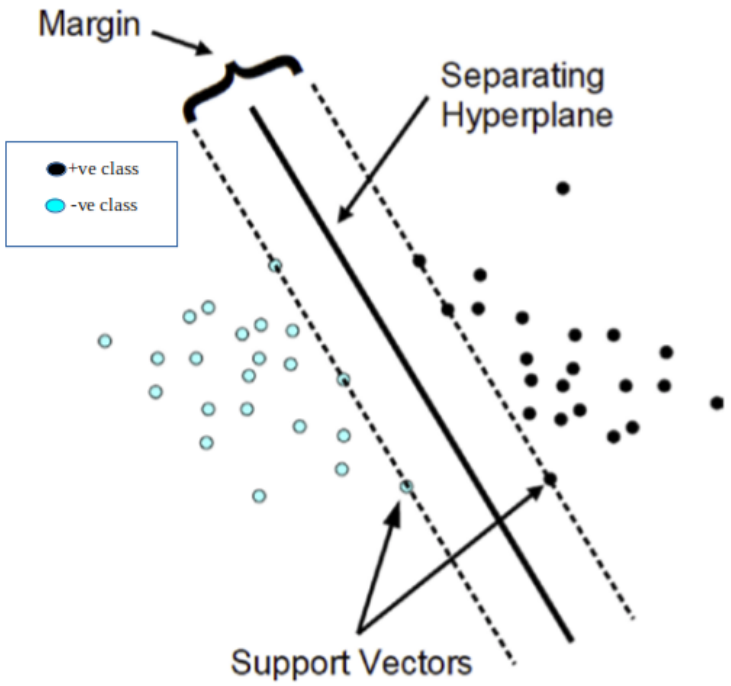

Figure 4.18: SVM classification

4.6.4 Confusion Matrix

The purpose of this study is to predict the popularity of tweets based on the volume of retweets. In order to meet that goal, we implement the supervised machine learning algorithm for constructing the model from the data. We examine the statistics that are calculated from a confusion matrix along with F-measure, precision, and recall metrics. The confusion matrix has a binary classification of the degree of retweeting that includes: (i) retweeting equal or more than 136 times which we refer to as popular tweet, and (ii) retweeting fewer than 136 times which we refer to as non-popular tweet. The following confusion matrix shows the information in more detail,

Actual Class	Predicted as popular tweet	Predicted as non-popular tweet
popular tweet	True popular tweet (TP)	False non-popular tweet (FN)
non-popular tweet	False popular tweet (FP)	True non-popular tweet (TN)

The parameters of this matrix are described as follows;

- **True popular tweets**: sample is belonging to the popular tweet class predicted as popular tweet
- **True non-popular tweets**: sample is belonging to the non-popular tweet class predicted as non-popular tweet
- **False popular tweet** : sample is belonging to the non-popular tweet class predicted as popular tweet
- **False non-popular tweet**: the sample is belonging to the popular tweet class predicted as non-popular tweet

The metrics for accuracy, recall, F1-score, and precision have a similar formula that we used in the previous chapter. Therefore, in the following section, we illustrate the results of our proposed model.

4.7 Results

4.7.1 Performance Analysis

In this section, we discuss our results by comparing the performance of three different classifiers, including random forest (RF), logistics regression (LR), and

support vector machine (SVM), by using topic analysis and vectorized features. In this study, the features are extracted by representing the tweet content into a matrix word where rows are the unique tweets and columns are the unique topics used in the corpus of tweet content. We used topic analysis by LDA method and n-grams by count vectorizer and TF-IDF vectorizer. In count vectorizer, we have uni-gram plus bi-gram, bi-grams, and tri-grams. The objective is having a matrix element that counts the frequency of the presence of words in a particular tweet. Thus, TF-IDF matrix calculates the term frequency and inverse document frequency of the word in the particular tweet. The following table shows the accuracy of each classifier by using different features.

Table 4.1: Performance evaluation of different features

	Accuracy (ACC)		
	Logistic Regression (LR)	**Support Vector Machine (SVM)**	**Random Forest (RF)**
Topics +Bi-grams TF-IDF vectorizer	**0.9519**	**0.9840**	**0.9903**
Topics LDA Modeling	**0.5754**	**0.5963**	**0.6040**
Uni-gram + Bi-grams TF-IDF vectorizer	**0.9386**	**0.9842**	**0.9899**
Bi-grams TF-IDF vectorizer	**0.9454**	**0.9654**	**0.9705**
Tri-grams TF-IDF vectorizer	**0.9234**	**0.9203**	**0.9300**

As it is shown on the table, the RF classifier performs better than the SVM and LR classifiers when topics (LDA) plus bi-grams TF-IDF vectorizer are used as features, but the accuracy is low when we only use topics by LDA model. Moreover, combining uni-grams and bi-grams with the TF-IDF vectorizer improves the accuracy of all three classifiers as the second method.

4.7.2 K-folds Cross-Validation

We apply the k-fold cross-validation (KVC) technique on all the classifiers on each category of features, by having k=5. This technique checks the possibility of over-fitting on our dataset and gives more accurate results. We explain the concept of this technique in previous chapter, thus, in this chapter, we only show the results of applying KVC on each classifier. Figure 4.19 shows the result of accuracy for RF classifier in each category of feature. The results show that topics plus bi-grams TF-IDF vectorizer and uni-gram plus bi-grams TF-IDF vectorizer have the highest accuracy for RF classifier.

Cross-Validation Accuracy for RF

	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average
Topics Analysis (LDA)	41.55%	61.26%	59.83%	58.82%	54.30%	55.15%
Topics plus Bi-grams TF-IDF	89.61%	99.32%	99.34%	99.27%	92.85%	96.08%
Uni-gram plus Bi-grams TF-IDF	89.67%	99.16%	99.15%	99.10%	93.78%	96.17%
Bi-grams TF-IDF	82.26%	95.06%	94.67%	94.39%	88.32%	90.94%
Tri-grams TF-IDF	75.65%	90.51%	90.73%	91.12%	84.51%	86.50%

Figure 4.19: KVC technique on RF classifier

Furthermore, Figure 4.20 shows the results of comparing the accuracy for each fold and the average of accuracy for SVM classifier. Figure 4.21 shows the results of the accuracy for different categories of feature for LR classifier by applying KVC technique.

The results for LR classifier shows that topics plus bi-grams TF-IDF vectorizer

Cross-validation Accuracy for SVM

	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average
Topics Analysis (LDA)	41.29%	60.68%	59.20%	56.42%	53.70%	54.26%
Topics plus Bi-grams TF-IDF	85.06%	98.92%	98.93%	99.00%	91.96%	94.77%
Uni-gram plus Bi-grams TF-IDF	86.11%	98.48%	98.80%	98.92%	92.29%	94.92%
Bi-grams TF-IDF	80.17%	95.22%	94.53%	94.56%	88.07%	90.51%
Tri-grams TF-IDF	75.61%	90.47%	90.77%	91.17%	84.52%	86.51%

Figure 4.20: KVC technique on SVM classifier

Cross-validation Accuracy for LR

	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5	Average
Topics Analysis (LDA)	41.48%	57.46%	59.81%	62.42%	58.44%	55.92%
Topics plus Bi-grams TF-IDF	78.53%	92.74%	92.11%	93.33%	85.80%	88.50%
Uni-gram plus Bi-grams TF-IDF	75.29%	91.21%	89.02%	89.01%	83.46%	85.60%
Bi-grams TF-IDF	76.36%	91.95%	90.43%	91.69%	85.90%	87.27%
Tri-grams TF-IDF	72.85%	89.56%	89.51%	90.44%	83.28%	85.13%

Figure 4.21: KVC technique on LR classifier

categories of features. Figure 4.22 shows the performance on the graph by comparing the accuracy after applying KCV.

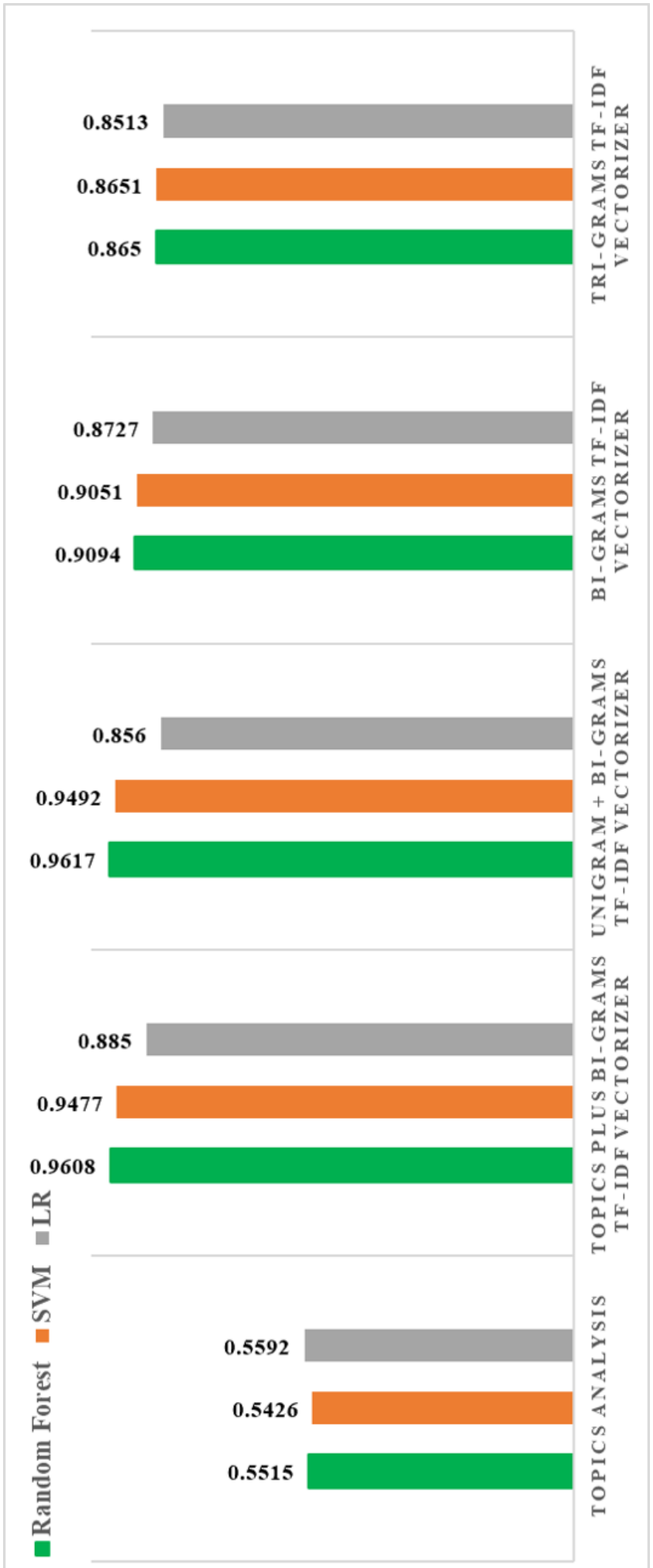

Figure 4.22: Performance Analysis for accuracy on bar graph

4.7.3 Receiver Operating Characteristics (ROC)

The receiver operating characteristics (ROC) curve is another essential task for measuring the performance of the classifiers. ROC is a probability curve and shows the strength of classifiers for distinguishing between classes, and thus, the higher ROC means better performance. The best model has the ROC near to the one, and the classier with poor performance has the ROC near to zero. The ROC curve shows the plot with the sensitivity of the true positive rate of retweets (TPR) against the false positive rate of retweets (FPR). The equation for the TPR and FPR are as follows:

$$TPR = \frac{TP}{TP + FN}$$

and

$$FPR = \frac{FP}{FP + TN}$$

The functionality of the ROC curves for RF classifier, LR classifier and SVM classifier by using different features with applying KVC are shown as follows. Figure 4.23 shows the ROC with applying KVC for using different features on RF classifier. As it is shown the Topics plus bi-grams TF-IDF vectorizer performance is near to the performance of applying uni-gram plus bi-grams TF-IDF vectorizer on RF classifier.

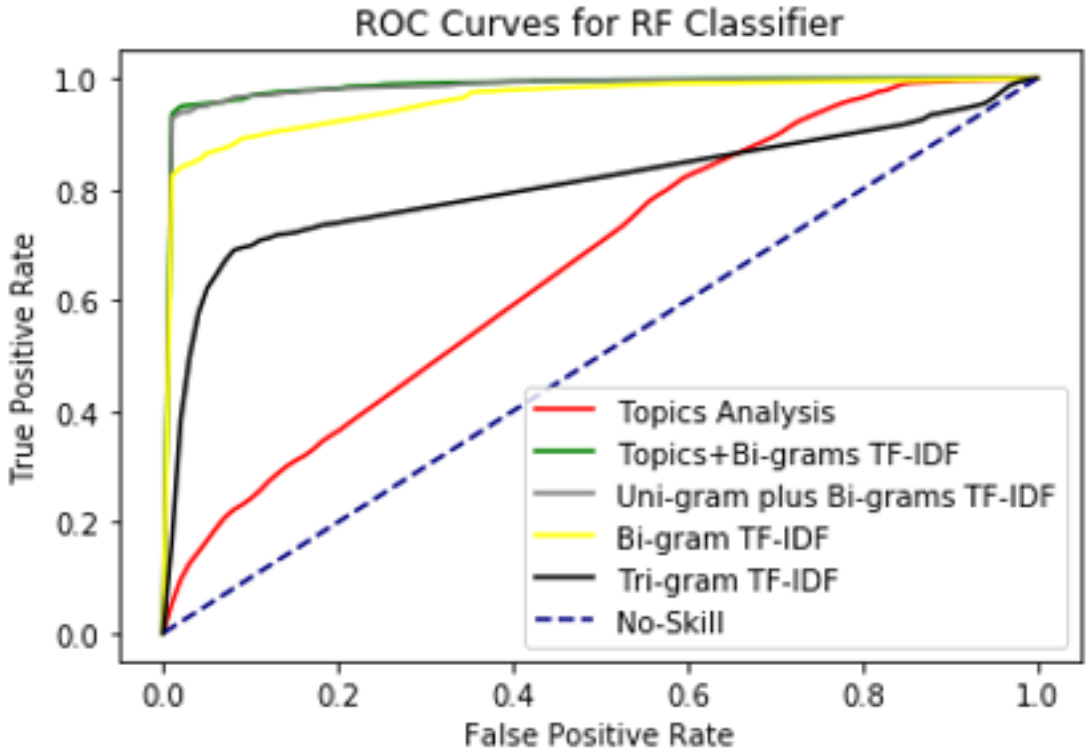

Figure 4.23: ROC with KVC of all the features for RF classifier

Furthermore, Figure 4.24 shows the ROC for using different features on SVM classifier, and Figure 4.25 shows the ROC for using different features on LR classifier. In all these classifiers topics plus bi-grams TF-IDF vectorizer as combined features enhance the performance of classifiers significantly compared to only using topic analysis.

Thus, in Figure 4.26 we compared the ROC of all three classifiers by applying combined features to identify which classifier has the better performance. Figure 4.26 shows RF classifier has the highest performance compared to LR and SVM classifiers.

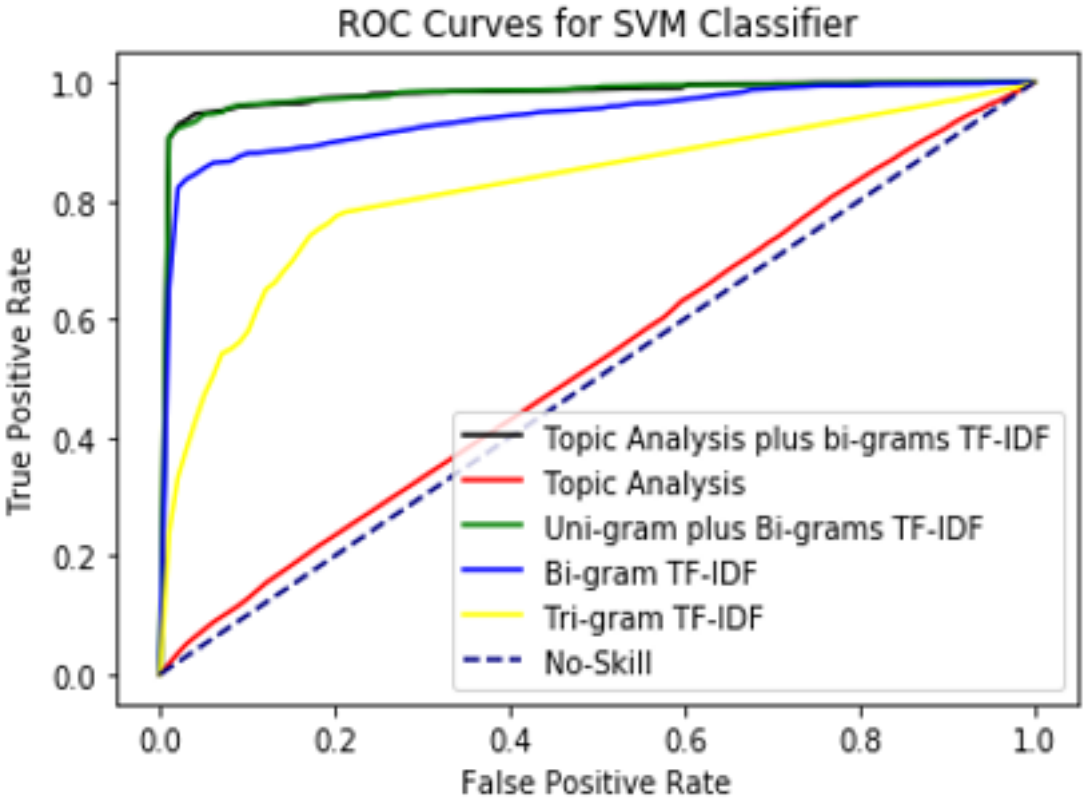

Figure 4.24: ROC with KVC technique of all features for SVM classifier

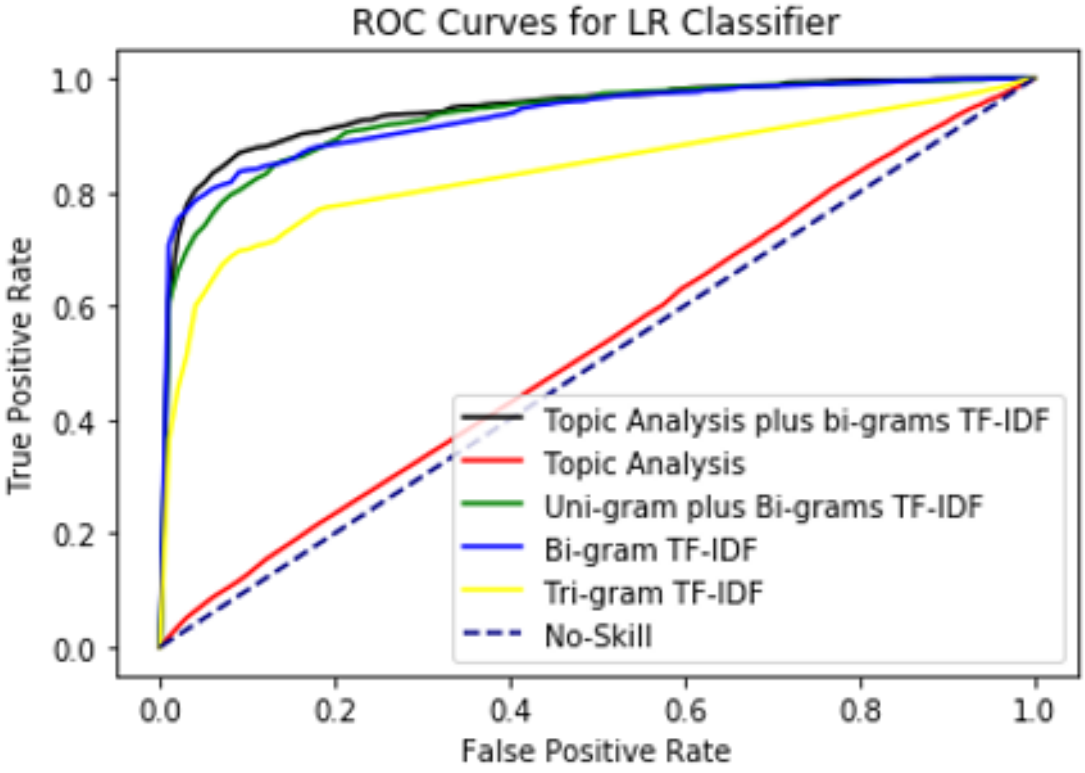

Figure 4.25: ROC with KVC of all features for LR classifier

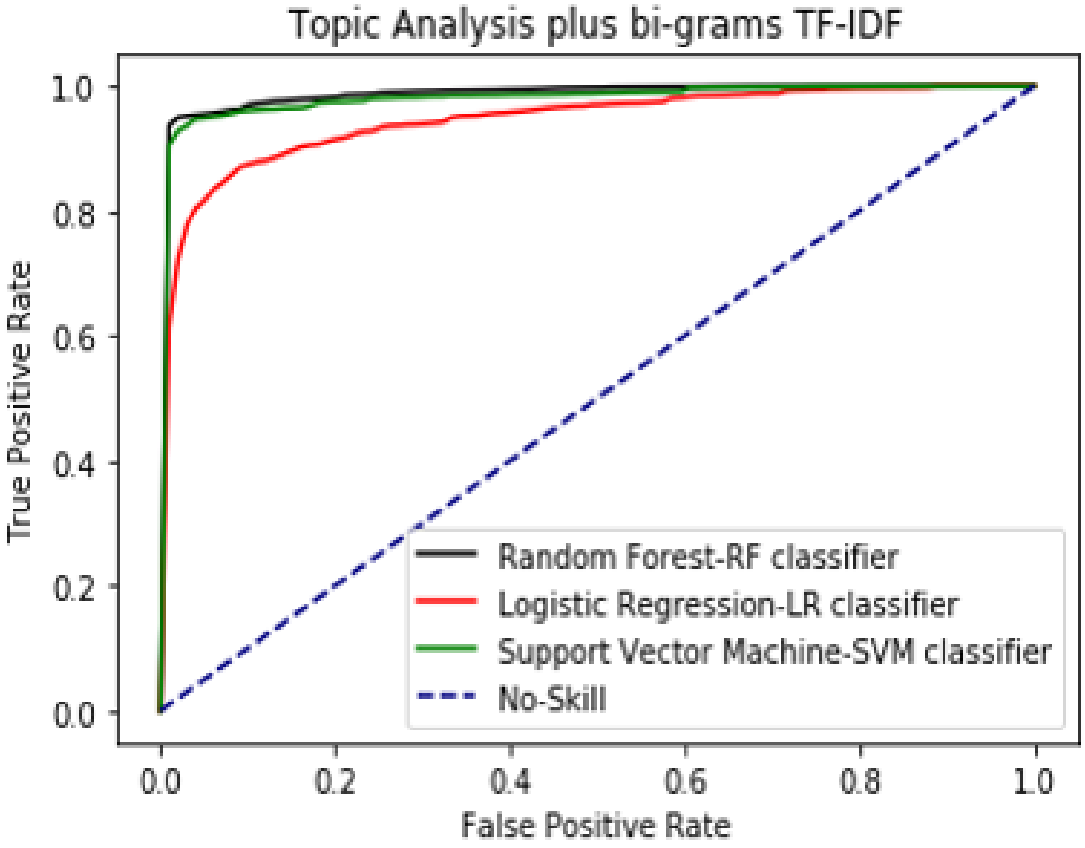

Figure 4.26: ROC with KVC technique for all the classifiers for Topics+Bi-grams TF-IDF vectorizer

4.8 Conclusion and Future Work

In this study, we analyze tweets about COVID-19 from March 27th, 2020 to June 5th, 2020, in the English language. Our Analysis related to the recent pandemic highlights that Twitter users in the United Stated are the most active user in the world on tweeting about COVID-19. People are more willing to retweet positive tweets (e.g., finding the cure, safe neighborhood, stay safe) compared to informative tweets (e.g., Wear mask, Stay home) and negative tweets(e.g., positive cases, end of the world), which enhances the risk of sharing misleading information. Furthermore, hashtags like Coronavirus and COVID-19 are still the most popular

hashtags in the world, even after several events including those arising from the Black Lives Matter movement.

We find that by analyzing the co-occurrence network of terms (bi-grams and tri-grams), we can use topics analysis by LDA technique for the short texts and solve the challenge related to it. We apply five different categories of features including (i) topics by using LDA model, (ii) n-grams by using TF-IDF vectorizer, which includes bi-grams, tri-grams, uni-grams plus bi-grams, and (iii) topics analysis plus bi-grams TF-IDF vectorizer on three supervised machine learning algorithms including random forest (RF) algorithm, support vector machine (SVM) algorithm, and logistic regression (LR) algorithm.

We aim to predict the popularity of tweets based on the volume of retweets by applying these classifiers on the tweets. We find that RF has the highest accuracy for predicting the popular tweets, and the performance of the algorithm improves by applying the topics plus bi-grams TF-IDF vectorizer. Furthermore, we find that uni-gram plus bi-grams TF-IDF vectorizer enhances the performance of classifiers close to accuracy of using topics plus bi-grams TF-IDF vectorizer. The reason might be related to having short texts with only 280 characters and also a sensitive subject like COVID-19. The result of this research can improve understanding of the users' preferences in retweeting content during the pandemic, and prevent the spread of misleading information by identifying these preferences at the early stage.

In future research, we can use the retweets frequency and the time of tweets to calculate the speed of spreading popular tweets based on their contents. In

line with this study, we can develop a recommendation system for a user who is tweeting with similar hashtags and keywords to find tweets with similar content that agree with or stand against the content of the user's tweets and so speed the retweeting process. Furthermore, we can apply the algorithm on a different dataset for multiple response variables. For instance, we can apply the algorithm on common COVID-19 symptoms to predict the severity of patient illnesses.

Bibliography

An, H. and Park, M. (2017). A study on the user perception in fashion design through social media text-mining. *Journal of the Korean Society of Clothing and Textiles*, 41(6):1060–1070.

Anthony, M. and Holden, S. B. (1998). Cross-validation for binary classification by real-valued functions: theoretical analysis. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 218–229.

Aseff, J. and Chade, H. (2008). An optimal auction with identity-dependent externalities. *The RAND Journal of Economics*, 39(3):731–746.

Bei, X., Gravin, N., Lu, P., and Tang, Z. G. (2019). Correlation-robust analysis of single item auction. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 193–208. SIAM.

Bengtsson, M. and Kock, S. (2000). *" Coopetition" in business Networks—to cooperate and compete simultaneously.*

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*, pages 1–10. IEEE.

Brandenburger, A. M. (1998). *Co-opetition*. Crown Business.

Breiman, L. (1999). 1 random forests–random features.

Budler, M. and Trkman, P. (2017). The role of game theory in the development of business models in supply chains. In *Technology & Engineering Management Conference (TEMSCON), 2017 IEEE*, pages 155–159. IEEE.

Bulow, J., Huang, M., and Klemperer, P. (1999). Toeholds and takeovers. *Journal of Political Economy*, 107(3):427–454.

Bulow, J. and Klemperer, P. (2002). Prices and the winner's curse. *RAND journal of Economics*, pages 1–21.

Cao, Q., Duan, W., and Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521.

Catal, C. and Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing*, 50:135–141.

Chen, Y.-J. and Vulcano, G. (2009). Effects of information disclosure under first-and second-price auctions in a supply chain setting. *Manufacturing & Service Operations Management*, 11(2):299–316.

Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Connors, L., Mudambi, S. M., and Schuff, D. (2011). Is it the review or the reviewer? a multi-method approach to determine the antecedents of online review helpfulness. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Dasgupta, P. and Maskin, E. (2000). Efficient auctions. *The Quarterly Journal of Economics*, 115(2):341–388.

Deneckere, R. and Peck, J. (1995). Competition over price and service rate when demand is stochastic: A strategic analysis. *The RAND Journal of Economics*, pages 148–162.

Dennison, J. A. and Montecchi, M. (2017). The effects of online consumer reviews on fashion clothing purchase intention: Peripheral cues and the moderating role of involvement. In *Advanced fashion technology and operations management*, pages 318–347. IGI Global.

El-Din, D. M. (2016). Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 7(1).

Ettinger, D. (2003). Bidding among friends and enemies.

Gardner, B. S. (2011). Responsive web design: Enriching the user experience. *Sigma Journal: Inside the Digital Ecosystem*, 11(1):13–19.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

Gnyawali, D. R., He, J., and Madhavan, R. (2006). Impact of co-opetition on firm competitive behavior: An empirical examination. *Journal of Management*, 32(4):507–530.

Gnyawali, D. R. and Madhavan, R. (2001). Cooperative networks and competitive dynamics: A structural embeddedness perspective. *Academy of Management review*, 26(3):431–445.

Gnyawali, D. R., Madhavan, R., He, J., and Bengtsson, M. (2016). The competition–cooperation paradox in inter-firm relationships: A conceptual framework. *Industrial Marketing Management*, 53:7–18.

Goeree, J. K. (2003). Bidding for the future: Signaling in auctions with an aftermarket. *Journal of Economic Theory*, 108(2):345–364.

Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence*, 31(4):343–397.

Granot, D. and Yin, S. (2008). Competition and cooperation in decentralized push and pull assembly systems. *Management Science*, 54(4):733–747.

Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58.

Hu, N., Bose, I., Gao, Y., and Liu, L. (2011a). Manipulation in digital word-of-mouth: A reality check for book reviews. *Decision Support Systems*, 50(3):627–635.

Hu, N., Bose, I., Koh, N. S., and Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 52(3):674–684.

Hu, N., Liu, L., and Sambamurthy, V. (2011b). Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3):614–626.

Huang, A. H., Chen, K., Yen, D. C., and Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Inderst, R. and Wey, C. (2004). The incentives for takeover in oligopoly. *International Journal of Industrial Organization*, 22(8-9):1067–1089.

Iyengar, J. J., Devcic, Z., Sproul, R. C., and Feeley, B. T. (2011). Nonoperative treatment of proximal humerus fractures: a systematic review. *Journal of orthopaedic trauma*, 25(10):612–617.

Jain, V. K. and Kumar, S. (2015). An effective approach to track levels of influenza-a (h1n1) pandemic in india using twitter. *Procedia Computer Science*, 70:801–807.

Jehiel, P. and Moldovanu, B. (1996). Strategic nonparticipation. *The RAND Journal of Economics*, pages 84–98.

Jehiel, P. and Moldovanu, B. (2000). Auctions with downstream interaction among buyers. *Rand journal of economics*, pages 768–791.

Jehiel, P. and Moldovanu, B. (2001). Efficient design with interdependent valuations. *Econometrica*, 69(5):1237–1259.

Jehiel, P., Moldovanu, B., and Stacchetti, E. (1999). Multidimensional mechanism design for auctions with externalities. *Journal of economic theory*, 85(2):258–293.

Ji, X., Chun, S. A., and Geller, J. (2013). Monitoring public health concerns using twitter sentiment classifications. In *2013 IEEE International Conference on Healthcare Informatics*, pages 335–344. IEEE.

Ji, X., Chun, S. A., Wei, Z., and Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5(1):13.

Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., and Luo, Z. (2017). R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*.

Kalaignanam, K., Shankar, V., and Varadarajan, R. (2007). Asymmetric new product development alliances: Win-win or win-lose partnerships? *Management Science*, 53(3):357–374.

Kanakaraj, M. and Guddeti, R. M. R. (2015). Nlp based sentiment analysis on twitter data using ensemble classifiers. In *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, pages 1–5. IEEE.

Kim, S., Kim, N., Pae, J. H., and Yip, L. (2013). Cooperate "and" compete: coopetition strategy in retailer-supplier relationships. *Journal of Business & Industrial Marketing*, 28(4):263–275.

Kolhe, N., Joshi, M., Jadhav, A., and Abhang, P. (2014). Fake reviewer groups' detection system. *Journal of Computer Engineering (IOSR-JCE)*, 16(1):6–9.

Korfiatis, N., GarcíA-Bariocanal, E., and SáNchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217.

Kostamis, D., Beil, D. R., and Duenyas, I. (2009). Total-cost procurement auctions: Impact of suppliers' cost adjustments on auction format choice. *Management Science*, 55(12):1985–1999.

Krishna, V. (2009). *Auction theory*. Academic press.

Krishna, V. and Maenner, E. (2001). Convex potentials with an application to mechanism design. *Econometrica*, 69(4):1113–1119.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Lado, A. A., Boyd, N. G., and Hanlon, S. C. (1997). Competition, cooperation, and the search for economic rents: A syncretic model. *Academy of management review*, 22(1):110–141.

Lazard, A. J., Scheinfeld, E., Bernhardt, J. M., Wilcox, G. B., and Suran, M. (2015). Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's ebola live twitter chat. *American journal of infection control*, 43(10):1109–1111.

Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948.

Liu, Y. and Liao, S. (2017). Granularity selection for cross-validation of svm. *Information Sciences*, 378:475–483.

Lorentziadis, P. L. (2014). Pricing in a supply chain for auction bidding under information asymmetry. *European Journal of Operational Research*, 237(3):871–886.

Lorentziadis, P. L. (2016). Optimal bidding in auctions from a game theory perspective. *European Journal of Operational Research*, 248(2):347–371.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., and Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1):87–103.

Macskassy, S. A. and Michelson, M. (2011). Why do people retweet? anti-homophily wins the day! In *Fifth International AAAI Conference on Weblogs and Social Media*.

Mamidi, R., Miller, M., Banerjee, T., Romine, W., and Sheth, A. (2019). Identifying key topics bearing negative sentiment on twitter: insights concerning the 2015-2016 zika epidemic. *JMIR Public Health and Surveillance*, 5(2):e11036.

Martineau, J. C. and Finin, T. (2009). Delta tf-idf: An improved feature space for sentiment analysis. In *Third international AAAI conference on weblogs and social media*.

Maskin, E. and Riley, J. (2000). Equilibrium in sealed high bid auctions. *The Review of Economic Studies*, 67(3):439–454.

McAfee, R. P. and McMillan, J. (1987). Auctions and bidding. *Journal of economic literature*, 25(2):699–738.

McMahon, C., Lowe, A., Culley, S., Corderoy, M., Crossland, R., Shah, T., and Stewart, D. (2004). Waypoint: an integrated search and retrieval system for engineering documents. *J. Comput. Inf. Sci. Eng.*, 4(4):329–338.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892.

Milgrom, P. R. and Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122.

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, 18(3):11–11.

Molnar, J. (2007). Pre-emptive horizontal mergers: theory and evidence.

Mudambi, S. M. and Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013). What yelp fake review filter might be doing? In *Seventh international AAAI conference on weblogs and social media*.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Nagarajan, M. and Sošić, G. (2007). Stable farsighted coalitions in competitive markets. *Management Science*, 53(1):29–45.

Nakayama, M. and Wan, Y. (2017). Exploratory study on anchoring: fake vote counts in consumer reviews affect judgments of information quality. *Journal of theoretical and applied electronic commerce research*, 12(1):1–20.

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference*, pages 1–7.

Norbäck, P.-J. and Persson, L. (2004). Privatization and foreign competition. *Journal of International Economics*, 62(2):409–416.

Odlum, M. and Yoon, S. (2015). What can we learn about the ebola outbreak from tweets? *American journal of infection control*, 43(6):563–571.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational

Pan, Y. and LIN, H.-f. (2008). Restaurant reviews mining based on semantic polarity analysis [j]. *Computer Engineering*, 17.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Papadimitriou, C. H. and Pierrakos, G. (2011). On optimal single-item auctions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 119–128.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Rane, A. and Kumar, A. (2018). Sentiment classification system of twitter data for us airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 769–773. IEEE.

Riley, J. G. and Samuelson, W. F. (1981). Optimal auctions. *The American Economic Review*, 71(3):381–392.

Roels, G., Karmarkar, U. S., and Carr, S. (2010). Contracting for collaborative services. *Management Science*, 56(5):849–863.

Romão, M. T., Moro, S., Rita, P., and Ramos, P. (2019). Leveraging a luxury fashion brand through social media. *European Research on Management and Business Economics*, 25(1):15–22.

Rothkopf, M. H. and Harstad, R. M. (1994). Modeling competitive bidding: A critical essay. *Management science*, 40(3):364–384.

Sain, S. R. (1996). The nature of statistical learning theory.

Salehan, M. and Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, 81:30–40.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Salvetti, F., Reichenbach, C., and Lewis, S. (2006). Opinion polarity identification of movie reviews. In *Computing attitude and affect in text: Theory and applications*, pages 303–316. Springer.

Sharma, K. and Lin, K.-I. (2013). Review spam detector with rating consistency check. In *Proceedings of the 51st ACM southeast conference*, pages 1–6.

Shrivastava, S. and Nair, P. S. (2015). Mood prediction on tweets using classification algorithm. *International Journal of Science and Research*, 4(11):295–299.

Singh, V. K., Piryani, R., Uddin, A., and Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717. IEEE.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE.

Szomszor, M., Kostkova, P., and St Louis, C. (2011). Twitter informatics: tracking and understanding public reaction during the 2009 swine flu pandemic. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 320–323. IEEE.

Tan, G. W. and Wei, K. K. (2006). An empirical study of web browsing behavior: Towards an effective website design. *Electronic Commerce Research and Applications*, 5(4):261–271.

Thongpapanl, N. and Ashraf, A. R. (2011). Enhancing online performance through website content and personalization. *Journal of computer information systems*, 52(1):3–13.

Varma, G. D. (2002). Standard auctions with identity-dependent externalities. *RAND Journal of Economics*, pages 689–708.

Varma, G. D. and Lopomo, G. (2010). Non-cooperative entry deterrence in license auctions: Dynamic versus sealed bid. *The Journal of Industrial Economics*, 58(2):450–476.

Wan, Y. and Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1318–1325. IEEE.

Wang, G., Xie, S., Liu, B., and Philip, S. Y. (2011). Review graph based online store review spammer detection. In *2011 IEEE 11th International Conference on Data Mining*, pages 1242–1247. IEEE.

Wang, M., Cao, D., Li, L., Li, S., and Ji, R. (2014). Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of international*

Wang, Z., Hou, T., Li, Z., and Song, D. (2015). Spotting fake reviewers using product review graph. *Journal of Computational Information Systems*, 11(16):5759–5767.

Willemsen, L. M., Neijens, P. C., Bronner, F., and De Ridder, J. A. (2011). "highly recommended!" the content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer-Mediated Communication*, 17(1):19–38.

Wilson, R. (1992). Strategic analysis of auctions. *Handbook of game theory with economic applications*, 1:227–279.

Wu, Z., Choi, T. Y., and Rungtusanatham, M. J. (2010). Supplier–supplier relationships in buyer–supplier–supplier triads: Implications for supplier performance. *Journal of Operations Management*, 28(2):115–123.

Xu, Y., Wu, X., and Wang, Q. (2014). Sentiment analysis of yelp's ratings based on text reviews.

Yami, S., Castaldo, S., Dagnino, B., and Le Roy, F. (2010). *Coopetition: winning strategies for the 21st century*. Edward Elgar Publishing.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., and Su, Z. (2010). Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1633–1636.

Ye, J. and Akoglu, L. (2015). Discovering opinion spammer groups by network footprints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 267–282. Springer.

Yu, B., Zhou, J., Zhang, Y., and Cao, Y. (2017). Identifying restaurant features via sentiment analysis on yelp reviews. *arXiv preprint arXiv:1709.08698*.

Zhang, J. and Wang, S. (2016). A fast leave-one-out cross-validation for svm-like family. *Neural Computing and Applications*, 27(6):1717–1730.

Zhang, Q., Gong, Y., Guo, Y., and Huang, X. (2015). Retweet behavior prediction using hierarchical dirichlet process. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.