Идеен подход за анализ на количествени данни от интервюта за целите на изследванията в областта на логистиката

A conceptual approach to analyzing quantitative interview data for logistics research

Николай Драгомиров[1]

Резюме

Несъмнено изследванията са важни както от изследователска гледна точка, така и за целите на управлението на бизнес единиците. Ако допреди години голяма част от проучванията се стремяха да се основават на анализа на количествени и силно структурирани данни, то във времето на информационен и технологичен бум се появяват и други потребности. В съвременните условия се налага използването все повече на количествени данни като изображения, аудио записи, видео, неструктурирани текстове и пр. Целта на доклада е да дефинира идеен подход за анализ количествени данни от текстове от проведени интервюта в областта на логистиката посредством използване на софтуерни системи. Неговите резултати могат да бъдат основа за провеждане на нови изследвания или разширяване на бъдещи такива. Неразделна част от доклада е представянето на някои конкретни софтуерни системи, които могат да бъдат използвани, както и на техните особености.

Abstract

Definitely, research is important both from a scientific point of view and for business management purposes. If years ago, much of the studies tended to be based on the analysis of quantitative and highly structured data, other needs are emerging in the days of information and technology boom. In today's environment, the use of quantitative data such as images, audio recordings, videos, unstructured texts, etc. is increasingly required. The aim of this paper is to define a conceptual approach for analysing quantitative data from texts of interviews conducted in the field of logistics by using software systems. Its results can be the basis for new studies or the extension of future ones. Some specific software systems that can be used and their features are an important part of the report.

JEL: C10, M21, C88

Увод

Допреди две десетилетия в областта на информационните системи, когато се говореше за данни, се възприемаше предимно съществуването на количествени такива. За качествените се споменаваше, но по-скоро имаха абстрактен характер. При съществуващите технологични решения обработката на изображения, видео, звук и пр. оставаше в зоната на футуризма. Днес обаче сме свидетели на това – почти всяко едно мобилно устройство с лекота разпознава изображения, обобщава разговори, обработва снимки и пр. Тези процеси няма как да не се пренесат и в областта на научните

[1] доц. д-р Николай Драгомиров, катедра „Логистика и вериги на доставките“ – УНСС, ORCID: 0000-0002-0923-962X, e-mail: ndragomirov@unwe.bg

изследвания и изследователският процес да бъде надграден с нови и лесно достъпни инструменти. Вече сме свидетели на концептуални модели за обработка на големи масиви от научна литература [1] – нещо, което преди би звучало доста футуристично и далечно.

Логистиката като наука има своето интензивно развитие в последните десетилетия и това е продиктувано от редица обективни причини. В основата на тези процеси стои конкурентоспособността и една специфична тенденция за изместването на нейния фокус от отделната фирма към веригата на доставките, към която принадлежи [2]. Така управлението на веригата на доставките като форма на еволюция на логистичната концепция се превръща в силно актуална тема. Всичко това довежда и до развитие в изследванията в областта и търсенето на нови и по-добри възможности.

Ако бъдат проследени проведените изследвания в областта, може да се отбележи, че в по-голямата си част разчитат на данни от анкети, които след това са обработени със статистически методи и/или в комбинация в провеждане на изследване на казуси. Донякъде това се дължи на причината, че повечето от изследванията имат по-фундаментален и общ характер. Навлизането на по-тесни проблеми на логистиката и управлението на веригата на доставките, както и процесите на дигитализация на света съответно довежда до търсенето на нови форми. Една от тях е свързана с възможността респондентите да отговарят свободно на дадени въпроси и след това резултатите да бъдат интерпретирани. Въпреки че като съдържание проблемът не е нещо ново, то в областта на логистиката има и своите специфики. Затова и целта на доклада е да дефинира идеен подход за анализ на количествени данни от текстове от проведени интервюта в областта на логистиката посредством използване на софтуерни системи. Неговите резултати могат да бъдат основа за провеждане на нови изследвания или разширяване на бъдещи такива. Неразделна част от доклада е представянето на някои конкретни софтуерни системи, които могат да бъдат използвани, както и на техните особености.

Възможности за анализ на текстови данни

Когато става дума за анализ на текстове, възможностите могат да бъдат класифицирани по различен начин. От една страна, е тяхната сложност, а от друга, са интерфейсът на работа и необходимостта от познания за програмиране на даден език. Казано по друг начин, едната алтернатива е използването на напълно завършени продукти с графичен интерфейс, които не изискват познания по програмиране и писане на код, докато при другите това е необходимо. С помощта на ИИ (изкуствен интелект) за събирането на конкретни алтернативи могат да се посочат решения като[1]:

- Отделни продукти – WordStat, MonkeyLearn, Lexalytics, RapidMiner, KNIME, IBM Watson NLP, Orage, Google Cloud Natural Language API и др.
- Библиотеки, които в повечето случаи са за Python – SpaCy, NLTK (Natural Language Toolkit), TextBlob, Gensim, Scikit-learn, PyTorch/NLP + TensorFlow, както и комбинации, включително и Hugging Face Transformers.

От посочените прави впечатление, че някои от тях имат различен фокус, но предлагат и функционалности за анализ на текст. Други, въпреки че са несочни към обработка на текстове, имат по-общи или по-специфични насоченост. Трети пък са специфични комбинации между библиотеки и платформи за машинно самообучение (machine learning platform) и пр. Като цяло възможностите са много и остава отворен въпросът за избора на конкретни решения.

[1] Списъците не са изчерпателни и имат само маркиращ характер на част от решенията.

В настоящия доклад са включени базови постановки, като са използвани две решения, а именно WordStat [3], както и NLTK [4]. Първият продукт е напълно завършен, предоставящ богат GUI (graphical user interface). Второто решение е в рамките на среда на Jupyter Lab [5], като са използвани библиотеките на NLTK и TextBlob в Python. Двете в комбинация са удачни за илюстриране на процеса по анализ на текстови данни и маркиране на някои особености. За целите на доклада са използвани междинни данни от проведени интервюта сред български логистични, търговски и производствени предприятия с илюстративен характер.

Подготовка за анализ и нейните особености в областта на логистиката

Сред основните стъпки в анализа се подрежда предварителната подготовка на текстовете, с която се цели тяхното адаптиране за конкретната система. Ако завършените продукти го правят автоматизирано или полуавтоматизирано с възможности за допълнителни настройки, то за останалите се изпълнява процес на токенизация (Tokenization). Този процес се определя в теорията и практиката като фундаментална стъпка за Natural Language Processing (NLP). По същество това е раздробяване на текста на по-малки части. Тези части могат да са думи, изречения или други текстови сегменти. Тяхното съществуване е предпоставка за последващата обработка на данните.

Пряко с процеса на токенизация се свързват два други процеса, а именно Stemming и Lemmatizition, отново причислявани към NLP. И при двата процеса се цели извеждане на общи части на думите, с което да се редуцира броят на токените. Например текстът “items are not ready for shipping” може да се трансформира посредством Stemming в [‘item’, ‘are’, ‘not’, ‘ready’, ‘for’, ‘ship’] и се вижда премахването на “s” и “ing”. При другия процес (Lemmatizition) се цели по-сложна морфологична трансформация, например извеждане на речниковите форми и пр. [6]

Съпътстваща стъпка е използването на речници, които могат да имат своите разновидности. Един от задължителните такива е речникът на думи, които да бъдат изключени. По същество това са думи и изрази, които се цели да бъдат премахнати от анализа, или т. нар stopwords. За всеки език съдържанието е различно и обикновено в началото на подготовката се използват универсални речници, които се надграждат по необходимост. В този и предходните процеси в областта на логистиката подобни речници могат да имат редица особености поради спецификата на терминологията. За съжаление в теорията и практиката се срещат сериозни разминавания при термини с едно и също значение. Подобни различия, например в областта на складирането, са породени от използването на различни термини в англоезичната и немскоезичната литература и така по същество няма смислово различие между тях [7]. Отделно ситуацията се усложнява с употребяването на вариации и специфики на фирмената култура и генерирането на собствени термини. Затова при тези обработки е важно да се възприеме единна терминология и текстовете да бъдат трансформирани, доколкото е възможно, преди да се продължи. Като цяло се доказва, че добре подготвените и изчистени бази данни подлежат на по-успешна обработка, включително при машинно самообучение [8], затова тази стъпка следва да не бъде подценявана. На този етап може да се посочи, че се разкрива потенциал за разработване на всякакъв вид заместващи или преводни речници в областта на логистиката и управлението на веригата на доставките, което от своя страна ще улесни подготовката на текстовете.

Друг съществен момент е възможността за изграждането на нови речници, които да бъдат по-конкретно използвани в последващите стъпки. Независимо от предходните етапи, на този може да се направи допълнително фокусиране върху определени теми. WordStat например позволява

изграждането на речници посредством автоматизирано извеждане на теми или чрез използване на фрази. В работната база данни, която се използва в разработката, присъстват свободни отговори на респондентите на зададените въпросни полета:

- Опишете взаимоотношенията на фирмата с нейните доставчици при изпълнение на логистичните дейности;
- Опишете взаимоотношенията на фирмата с нейните клиенти при изпълнение на логистичните дейности;
- Опишете основните предизвикателства пред фирмената логистика.

При опит за извеждането на дванадесет теми с използване на Factor analysis (без да се конкретизират настройките), другата възможност е Non-Negative Matrix Factorization – NMF, могат да се посочат част от резултатите в Таблица 1[1].

Таблица 1: Част от изведените теми от софтуера

TOPIC	KEYWORDS
КЛИЕНТИТЕ ОБСЛУЖВАНЕ	КЛИЕНТИТЕ; ОБСЛУЖВАНЕ; КЛИЕНТА; БЪРЗО; ВРЪЗКА; ПРАТКИТЕ; ПОРЪЧКА; ПРЕДЛАГА; ПРЕДОСТАВЯ; ИНФОРМАЦИЯ; ПОЛУЧАВА; ПРОЦЕС; ПОРЪЧКИТЕ; ИЗПЪЛНЕНИЕ; ДОСТАВКА
ВИСОКО КАЧЕСТВО	ВИСОКО; КАЧЕСТВО; ПОДДЪРЖА; ОБСЛУЖВАНЕ; СТРАНИ; ОСИГУРЯВА; ИЗИСКВАНИЯ; ВЗАИМООТНОШЕНИЯТА; КОМУНИКАЦИЯ
СВОИТЕ УСЛУГИ	СВОИТЕ; УСЛУГИ; КОМПАНИЯТА; ПРЕДЛАГА; ПРЕДОСТАВЯ; КЛИЕНТИТЕ; ПРОДУКТИ; ЛОГИСТИЧНИТЕ; РАБОТИ; ОТНОШЕНИЯ; ОСИГУРЯВА; КАЧЕСТВО; ВЗАИМООТНОШЕНИЯТА; ДЕЙНОСТИ; РАЗЛИЧНИ; ОБСЛУЖВАНЕ; ГОЛЯМА; ДОСТАВЧИЦИТЕ; ФИРМАТА; ПОДДЪРЖА; ЛОГИСТИЧНИ
СКЛАД	СКЛАДА; СТОКАТА; СТОКИ; КОЛИЧЕСТВО; ДОСТАВЧИК; ПОРЪЧКА; ПОРЪЧКИТЕ; ДОСТАВКА; ГОЛЯМА
КОМУНИКАЦИЯ ИНФОРМАЦИЯ	КОМУНИКАЦИЯ; ИНФОРМАЦИЯ; ПОДДЪРЖА; ПРОБЛЕМИ; ПРЕДОСТАВЯ; ОСИГУРЯВА; ДОСТАВЧИЦИТЕ; ВЗАИМООТНОШЕНИЯТА; ПРАТКИТЕ; НЕОБХОДИМИТЕ; ФИРМАТА; ВКЛЮЧВА; ДОСТАВКА; ПОРЪЧКИТЕ; ЕФЕКТИВНО
ЕФЕКТИВНО УПРАВЛЕНИЕ	ЕФЕКТИВНО; УПРАВЛЕНИЕ; ПРОЦЕС; ВКЛЮЧВА; ЛОГИСТИЧНИ; РАЗЛИЧНИ; ИЗПОЛЗВА; РАЗХОДИТЕ; ОСИГУРЯВА; ЛОГИСТИЧНИТЕ; ИЗИСКВАНИЯ
УСЛОВИЯ ЦЕНИ	УСЛОВИЯ; ЦЕНИ; ДОГОВОРИ; КАЧЕСТВО; ОПРЕДЕЛЕН; ВКЛЮЧВА; ИЗИСКВАНИЯ; ОТНОШЕНИЯ; ДОСТАВКА; ФИРМАТА; ПРЕДЛАГА; ПРЕДОСТАВЯ; ИНФОРМАЦИЯ; ДОСТАВЧИЦИТЕ; РАЗХОДИТЕ
ДОГОВОРИ СТРАНИ	ДОГОВОРИ; СТРАНИ; РАБОТИ; ФИРМА; ДОСТАВЧИК; ОТНОШЕНИЯ; ВЗАИМООТНОШЕНИЯТА; ФИРМАТА; ДОСТАВЧИЦИТЕ; КОМУНИКАЦИЯ
ТРАНСПОРТ	ТРАНСПОРТ; ИЗВЪРШВА; ФИРМА; ИЗПОЛЗВА; УСЛУГИ; РАЗХОДИТЕ; КЛИЕНТА; ДОСТАВКА; ГОЛЯМА; ЛОГИСТИЧНИТЕ; МАТЕРИАЛИ

Source: Edited software output

Така дефинираните теми или част от тях могат да бъдат използвани за дефинирането на речник и продължаването на анализа. Сходна е ситуацията с използването на фрази. След определянето на

[1] Имената на някои от темите са редактирани с цел по-добро разбиране на значението. Например в оригинал „Склад“ е дефинирана като „Склада стоката“. Също така използваната база данни е в процес на изграждане и използването ѝ има чисто демонстративна цел.

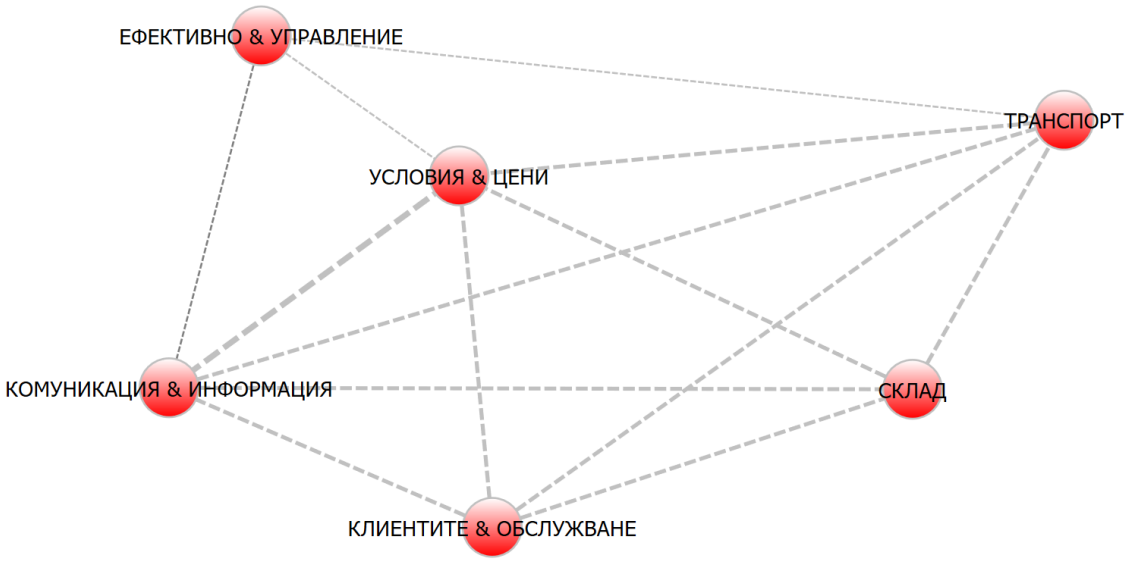
тяхната честота е възможно част от тях да бъдат включени в речници за категоризация. Може да се отбележи, че най-често е използването на bigrams и trigrams (n-grams), съответно фрази с две и три думи.

Развитие на анализа

След подготовката на базата данни и изграждането на речниците е възможно анализът да бъде продължен в редица посоки. Независимо от платформата, която се използва, почти задължителни са извеждането на разни статистики за текста. Най-често под формата на bag of words, wordcloud или друга алтернативата в табличен или графичен вид. На този етап е удачно да се правят различни видове класификации, утвърдени в изследователската практика. В областта на проучванията по логистика е удачно те да бъдат според:

- Участници в логистичните процеси – търговски, производствени или логистични фирми;
- Размер на организацията – микро и малки предприятия, средни и големи предприятия.

Друг важен метод за анализ е използването на различни видове класификационни подходи и обследване на връзките между отделните резултативни елементи. Един път създадени, речниците дават възможност елементите да бъдат йерархично класифицирани и след това да се направи анализ на връзките между тях. При използване на речника, посочен в таблица 1, могат да се определят няколко негови записа и те да бъдат изследвани по-задълбочено. Графично примерната връзка между тях е посочена на фиг. 1.



Source: Software output

Фигура 1: Връзка между няколко елемента от речника

В случая са избрани част от записите и силата на връзките е филтрирана, за да се избегнат шумовете. От данните ясно проличава наличието на силна връзка около обслужване на клиенти, комуникация и условия/цени. Тази група съответно има по-силна връзка с транспорта, отколкото със склада. Отделно от това ефективното управление е насочено по-скоро към комуникацията и информацията,

отколкото към останалите елементи. Анализът може да се продължи и по този начин да бъдат дефинирани редица изследователски измерения и вероятно най-важното се състои в това да се обследват връзките между тях.

Могат да бъдат посочени и редица други методи за анализ, като например[1]:

- Групиране на думи и съпоставка спрямо останалите, включително прилагане на дескриптивни статистически методи. Представяне на резултатите във векторно пространство;
- Честотни разпределения по групи – видове организации, размер, време и пр. Включително търсене на нови клъстери по тези групи;
- Едновременна поява/присъствие на думи;
- Обследване на записите на речника в какъв контекст са използвани. Най-често се провежда анализ на предходните и последващите думи в текста;
- Идентифициране на частите на речта (Part-of-speech tagging – POS);
- Оценка на цели изречения според това дали са позитивни, негативни или неутрални по даден въпрос (Average sentiment score);
- Степен на разнообразие на използваните думи и богатството на речника;
- Степен на сходство;
- Други.

Някои автори посочват и други методи, като включват и големи категории с редица методи за “Dimension Reduction”, отделно и за “Logistic Regression” и пр. [9], което позволява да се обогати значително списъкът при необходимост. При провеждането на подобни анализи от голямо значение е езикът и наличието на ресурси за анализ. Като цяло може да се посочи, че за английския език са налични най-много решения, докато за българския език съществуват множество бариери [10]. Ето защо анализът преди всичко трябва да се съобрази и с наличните ресурси под формата на софтуерни функционалности и програмни библиотеки към дадения момент.

Основни изводи

Развитието на изследователските подходи е факт, като причината за това може да се търси в редица посоки, включително и в процесите на дигитализация и дигитална трансформация на световно равнище. Ако досега повечето проучвания са били насочени към анализ на количествени и силно структурирани данни, то сега се появяват и нови потребности. Използването на количествени данни става все по-необходимо и изследванията в областта на логистиката следва да не изостават. Затова в настоящата разработка е предложен базов концептуален подход за използване на софтуерни системи за анализ на количествени данни от текстове на интервюта, направени в областта на логистиката. Определено съвременните софтуерни системи имат достатъчен потенциал за това и свободата на изследователите е доста висока. При тяхното използване могат да се прилагат както универсални, така и по-специализирани подходи според конкретните изследователски цели. Важен момент в подобен вид изследвания е доброто познаване на отделните методи и при необходимост уеднаквяването на използваната терминология. Отделно от това при подобни разработки би следвало

[1] Списъкът е изведен на база функционалности на различни видове решения за текстови анализ, като някои от тях имат ограничено приложение на български език.

да се прилагат различни видове подходи, които са утвърдени в предходни изследвания, посветени на проблемите на логистиката и управлението на веригите на доставките.

References

[1] N. Dragomirov, "Digital tools usage for literature review," in *Logistics in times of crisis: Challenges and solutions*, Varna, 2022, pp. 158–162. doi: 10.5281/zenodo.7324287.

[2] M. Christopher, *Logistics & supply chain management*, 4. ed. Harlow: Financial Times Prentice Hall, 2011.

[3] Provalis Research, *WordStat*. Accessed: Sep. 13, 2024. [Online]. Available: https://provalisresearch.com

[4] NLTK Project, *NLTK (Natural Language Toolkit)*. [Online]. Available: https://www.nltk.org

[5] Project Jupyter, *JupyterLab*. [Online]. Available: https://jupyter.org

[6] Jacob Murel and Eda Kavlakoglu, "What are stemming and lemmatization?" Accessed: Sep. 17, 2024. [Online]. Available: https://www.ibm.com/topics/stemming-lemmatization

[7] Н. Драгомиров, М. Воденичарова, М. Стефанов, Л. Михова, М. Кътева, and А. Метиева, *Складови системи в логистиката - управленски практики и тенденции*. ИК - УНСС, 2022.

[8] F. González, M. Torres-Ruiz, G. Rivera-Torruco, L. Chonona-Hernández, and R. Quintero, "A Natural-Language-Processing-Based Method for the Clustering and Analysis of Movie Reviews and Classification by Genre," *Mathematics*, vol. 11, no. 23, p. 4735, Nov. 2023, doi: 10.3390/math11234735.

[9] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text Mining in Big Data Analytics," *BDCC*, vol. 4, no. 1, p. 1, Jan. 2020, doi: 10.3390/bdcc4010001.

[10] G. Hristova, "Text Analytics in Bulgarian: An Overview and Future Directions," *Cybernetics and Information Technologies*, vol. 21, no. 3, pp. 3–23, Sep. 2021, doi: 10.2478/cait-2021-0027.