Uzbek Texts Sentiment Analysis: Database Development

Saboxat Allanazarova and Dilrabo Elova
Tashkent State University of Uzbek Language and Literature, Tashkent, Uzbekistan

Keywords: Sentiment Analysis, Social Network, Reviews, Customer's Opinion, Emotional Coloring.

Abstract: In the digital era, research centered on exploring customer sentiments towards service quality, brands, products, and events predominates. Social networks are pivotal platforms for this investigation, leveraging online reviews to gauge consumer opinions across diverse domains. Among the myriad approaches to sentiment analysis, this article focuses on the lexicon-based method, employing a dictionary of emotive words tailored to the Uzbek language. This method involves parsing text to identify words with emotional connotations, enabling a nuanced understanding of customer sentiment. By harnessing lexicon-based sentiment analysis, researchers can dissect and interpret consumer perceptions, providing invaluable insights for businesses and organisations seeking to enhance customer satisfaction and refine their offerings.

1 INTRODUCTION

As a result of the rapid advancements in science and technology, the influx of information via social media platforms has soared exponentially. One of the primary challenges confronting users is the scarcity of time. Given the sheer volume of available data, it becomes impractical to digest every piece of information. Consequently, there's an escalating demand for methods to sift through this deluge of data.

This surge in interest has fuelled exploration in fields such as Natural Language Processing (NLP), Machine Learning, Data Science, and Artificial Intelligence. In the 21st century, marked by technological prowess and intellectual evolution, a novel approach to learning has emerged (Liu 2012, Medhat et al 2014).

The Uzbek language, considered one of the world's developed languages, suffers from underutilization within the realm of information technology. Regrettably, our language's inherent potential remains largely untapped. Particularly in areas like Sentiment Analysis (commonly known as opinion mining), the process of data collection and processing proves to be excessively time-consuming for Uzbek. This highlights the pressing need for further development and integration of Uzbek language capabilities into modern technological systems. Efforts to enhance the efficiency and effectiveness of language processing in Uzbek are vital to harnessing its full potential in the digital age (Divyapushpalakshmi et al 2021, Ahmedova et al 2021).

2 RESEARCH METHODOLOGY

The research methodology for sentiment analysis outlined above integrates both machine learning and lexicon-based approaches.

The lexicon-based method relies on semantic analysis of lexical units to gauge sentiment polarity within text, considering factors such as the speaker's attitude, emotional state, and the context of speech.

By assessing the semantic orientation of words and sentences, this approach quantifies subjectivity and opinion. Emotionally charged language is pivotal in sentiment analysis as it conveys the speaker's feelings and attitudes towards a given subject.

The methodology recognises the significance of informal comments prevalent in social media platforms like Twitter and Facebook, as well as review sites such as the Google Play Store, where users express their opinions about products and services.

However, the methodology acknowledges the noise inherent in such opinionated data sources and emphasizes the need for robust techniques to distil meaningful insights.

Allanazarova, S. and Elova, D.
Uzbek Texts Sentiment Analysis: Database Development.
DOI: 10.5220/0012915100003882

In *Proceedings of the 2nd Pamir Transboundary Conference for Sustainable Societies (PAMIR-2 2023)*, pages 744-747
ISBN: 978-989-758-723-8

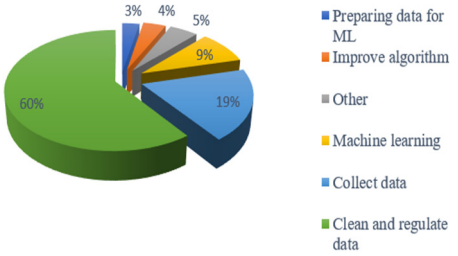

Figure 1: Sentiment Analysis Essentials: From Data Collection to Processing.

Emphasis is placed on the importance of extracting valuable information from opinionated sources despite the inherent noise.

The methodology underscores the relevance of sentiments expressed by users in social media comments and reviews, as they offer authentic reactions, personal experiences, and recognitions. These sentiments play a pivotal role in influencing decisions of potential consumers when evaluating businesses or services.

Therefore, the methodology highlights the necessity of employing advanced sentiment analysis techniques to sift through noisy data sources effectively. By leveraging both machine learning and lexicon-based approaches, the methodology aims to provide accurate sentiment classification, thereby facilitating informed decision-making processes based on user-generated content.

3 RESULTS AND DISCUSSION

Our goal is to build the first labeled dataset for sentiment analysis in Uzbek language texts, obtained from the "Annotated dictionary of the Uzbek language." This dictionary is a five-volume annotated dictionary of the Uzbek language and one of the most comprehensive Uzbek annotated dictionaries published to date. The dictionary is based on a two-volume annotated dictionary published in 1981. Furthermore, we have built a larger manually dataset of emotional coloring words and idioms. This annotated dataset contains three categories: positive, negative, and neutral. As in all languages, the Uzbek language also has methods and tools that create expressiveness and emotionality, and ways of using it. Phonetic, lexical-phraseological, morphological, and syntactic methods of expressive-emotional expression in the Uzbek language are given. The lexicon of the language has great potential for conveying information with the subtlest meaning and stylistic coloring. The stylistic features of units in the vocabulary of the language are studied in the course. Among them, especially synonyms and antonyms are tools rich in stylistic possibilities. Words with stylistic color are especially expressed in the synonymous line. A synonym string can consist of only two words or several words.

Figure 2: Synonymous Synsets in Uzbek: A Multi-line Example.

From this synonymous group, Yuz is the main, dominant word, which expresses a neutral assessment and does not choose speech styles. The words Aft, Turq, Bashara, Bet are negative words. Words such as Chehra, Diydor, Oraz, Siymo have a positive connotation. Each word in this synonymous line is used selectively in different speech situations. Most of the words expressing an opinion are adjectives. For example: He is knowledgeable. His conversations are interesting. The adjectives "knowledgeable" and "interesting" in this sentence are positive words, and the nouns traffic, sad, and unemployment express a negative meaning. Phrases show the strong influence of events, symbols on the human mind, the full expression of the result of this strong influence in speech, in general, expresses the influence of thought at a high level.

The phenomenon of gradation, which is present in expressions, allows expressing expressiveness to the extent necessary. For example, if the phrase "Sochi tikka bo‘lmoq" is considered the lowest level in terms of expressiveness, it reinforces as follows using certain phonetic or grammatical means and creates a unique gradation.

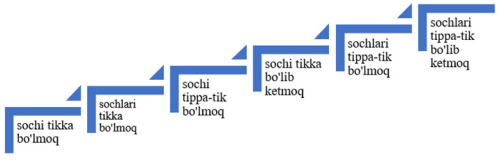

Figure 3: Graduonymic Hierarchy: Phrases in Uzbek Language.

Lexical synonyms are distinguished by the gradation of the meaning of the term, stylistic characteristics, sometimes having a negative-positive color. In the following years, graduonyms were studied in depth by the scientists of our republic. Graduonymy is a ranking of synonyms based on a certain difference, based on a chain connection.

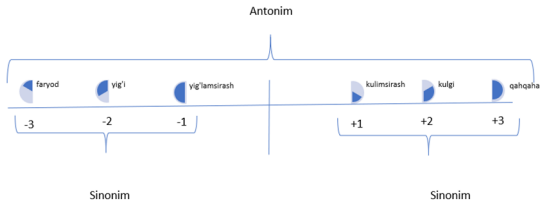

Figure 4: Lexical Poles: Synonyms and Antonyms in Perspective.

Although sentiment analysis is a very active area of research today, a number of complex problems remain in this field. First, the problem of sarcasm; Sarcasm is a complex form of speech units in which the opposite of what the commenter or writer intended is said or written. In sentiment analysis, sarcasm and ironic sentences are very difficult to analyze because the idea is not clearly and directly expressed.

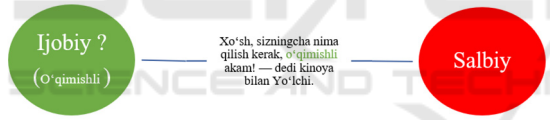

Figure 5: Irony in Uzbek: Examples of Sarcastic Expression.

Even though positive words are included in such sentences, the negative meaning is understood through the tone of speech. Sarcasm and ironic words are used with specific, different intonation in oral speech, but in writing, they are often used with quotation marks[9]. The computer does not understand the opposite meaning in this sentence. As a result, it evaluates the opinion as positive. Often, words that express a positive assessment contain a negative meaning. For example: Today, the geography teacher fell and broke his leg. After this "good news," the whole class was empty.

Polysemantisms are neutral with one meaning (nominative meanings) and serve to express an expressive-emotional thought with another meaning (figuratively meaning):

1. Hayvon (animal) – objective word – neutral meaning;
2. Hayvon (beast) – abusive word – negative meaning.

Analysis factors and mathematical models are necessary to create a system for semantic analysis of polysemantic words [10]. Polysemantic and homonymous words are close words. That is why the issue of studying the phenomena of homonymy and polysemy has not lost its relevance until now. According to world experience, polysemantic words are semantically analyzed based on statistical calculations and probability theory on a large amount of contextual data [11]. Researcher Sh. Gulyamova[12] expressed her opinion on the semantic analysis of polysemous words in the Uzbek language. It is recommended to use Markov models in the semantic analysis of polysemantic words in the Uzbek language.

In linguistics, expressions such as extremely negative attitude, discrimination, disdain, and insult are very clearly visible in insulting words called vulgarisms[13]. Such words are expressed in speech not according to their nominative meaning but according to their connotative meaning. Insulting words are used in the speech of heroes in literary works and in comments on social networks[14]. Vulgarisms also differ in terms of gender; that is, they are used differently in the speech of men and women.

1. Vulgarisms used for women: g'ar, megajin, manjalaqi… «Otasidan hovliyu mashinadan boshqa yana nimalar qoldi ekan?» – deb atrofingda hid olib yurgan bitta shu megajin emasdir!
2. Vulgarisms used for men: takasaltang, landovur, nomard… – Qo'lga tushding-ku, qani endi xo'jayinga javob berib ko'r, ha nomard! As we have seen in the above examples, the accuracy of sentiment analysis is increased by including insults in the linguistic database.

There are sentences that are often used in the Uzbek people's culture of communication, in which the persons who have entered into a relationship wish each other well[15]. Such sentences are said when praying, congratulating on a relationship, farewell, greeting, condolence (in native language: duo, olqish, qarg'ish) [16]. Such sentences have stabilized semantically and formally in the Uzbek language, most of them have become speech etiquette, and all of them have a positive meaning.

Table 1: Uzbek Language Dynamics: Communication Styles.

Subjectivity	Example:	Classification:
Olqish	Aylanay, o‘rgulay, jonim tasadduq bo‘lsin!	positive
Qarg‘ish	Yonimga hech kelmasin, baloga uchrasin!	negative
Duo	“Qo‘l-ko‘zing dard ko‘rmasin iloyim”, - dedi xola mehr bilan.	positive

Uzbek, a Turkic language, stands as the primary official and sole designated national language of Uzbekistan. This language, spoken by Uzbeks, follows a null-subject, agglutinative structure and boasts a multitude of dialects, each varying significantly across regions, thus presenting complex challenges. While languages within the Turkic family, such as Turkish and Kazakh, have made notable strides in sentiment analysis, determining the number of emotional words in Uzbek proves to be a formidable task. This endeavor demands robust linguistic expertise and extensive vocabulary research. One initial hurdle lies in the absence of a linguistic database tailored for Uzbek, hindering sentiment analysis efforts.

4 CONCLUSION

In conclusion, the development of a linguistic database for sentiment analysis in Uzbek texts necessitates meticulous consideration of the language's unique features. While the chosen approach promises insight into emotional expression, challenges arise regarding the time-consuming task of compiling synonyms and antonyms. Moreover, the prevalence of informal lexicon in platforms like Twitter underscores the need to accommodate informal language within the database.

Addressing these challenges, the decision to incorporate informal lexicon, including insults and emoticons, proves crucial. This expansion not only enables the classification of formal texts but also empowers the database to analyze informal communication effectively. By bridging the gap between formal and informal language usage, the database enhances its utility in understanding sentiment across various linguistic contexts.

REFERENCES

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

Abdullayev A. (1976). The expressiveness characteristics of phraseological units. Uzbek language and literature, 5, 36-39.

Makhmudova N. (2021). Functional-semantic field of graduation category in the English and Uzbek languages. Philology Matters, 2021(4), Article 4.

Ahmedova X. (2021). Mathematical models that distinguish homonymy in the framework of a word series. Electronic journal of actual problems of modern science, Education and training, October, 2021-10/1. ISSN 2181-9750.

Divyapushpalakshmi M, Ramalakshmi R. (2021). An efficient sentimental analysis using hybrid deep learning and optimization technique for Twitter using parts of speech (POS) tagging. International Journal of Speech Technology, 24(2), 329–339.