A SYSTEM FOR RETRIEVAL, STORAGE, AND PHONEMIC REPRESENTATION OF INTERNET NEWS FOR STUDYING THE RELATIONSHIP BETWEEN THE PHONEMIC CONTENT AND THE EMOTIONAL VALENCE OF WRITTEN SOURCES

Rosen Stefanov, Martin Konstantinov, Velina Slavova

Abstract: *The computer system presented here was developed in order to collect data allowing testing statistically the hypothesis that the sublexical level of language carries information about the emotional valence (positive - negative emotion) expressed in a text. The system collects data from authoritative news sources, stores them, represents them at the sublexical level as meaning-unloaded word parts, and analyzes them in order to investigate the relationship between phonetic content and emotional valence of written sources. Given the linguistic dependency of the study, based on dependencies derived from the analysis of an emotionally rated corpus of English, the test news items that the presented system accumulates are from native English speaking authors.*

Keywords*: Text processing, Web system, Text emotion recognition, Emotional valence.*

ITHEA Keywords: H.3.1 Content Analysis and Indexing**,** H.3.3 Information Search and Retrieval

DOI*: https://doi.org/10.54521/ijita30-03-p02*

Introduction

The study of emotions has intensified recently, especially in relation to machine recognition. Research indicates that emotions impact various levels — physiological, mental, and cognitive. Currently, two main models of emotions are employed. The discrete model categorizes emotions with specific names (joy, sadness, anger, etc.), with variations in number and composition based on researchers. The second model, VAD (Valence-Arousal-Dominance), is continuous, representing emotional states on a spatial scale, often along two axes: Valence, indicating positivity or negativity, and Arousal, indicating the level of mental arousal associated with the emotion. The system described here relies on the second model due to its versatility and the abundance of evidence supporting its benefits.

Studies on brain activity reveal distinct patterns associated with emotional states in the cerebral cortex. For instance, certain areas of the cortex exhibit responses exclusively to negative words, highlighting the non-linguistic involvement in emotion processing. Particularly pertinent to the approach outlined is the finding that when reading text, the auditory cortex (linked with auditory processing) becomes automatically engaged, processing written text in a manner akin to spoken language input.

Studies across various languages in both poetry and prose have highlighted the relationship between emotion and language. Linguists and psycholinguists have identified connections between individual emotions and specific phonemes. Recent research in the field of "sound symbolism" has explored the link between phonetic composition and word meaning. While these relationships are

not directly observable, statistical methods applied to linguistic data have aided in their discovery.

A study by Kawahara and Shinohara (2012) revealed that vowel-consonant pairs presented as acoustic stimuli evoked both a sense of magnitude and emotion in subjects, suggesting that sublexical components of speech convey emotional content through their phonetic characteristics '[Kawahara & Shinohara (2012)]. This discovery led to the hypothesis that sublexical components of speech, which are parts of words, convey emotional content through their phonetic properties. This laid to apply a Gestalt approach to text, which views text as a unified whole consisting of interconnected levels — sentences, words, and word parts — each contributing to encoding emotional valence [Slavova, 2021]. The application of the Gestalt approach to text in recent years has yielded significant insights into the relationship between text and emotion. For instance, a series of corpus studies conducted on the German language have demonstrated that all levels of language, including syllables, convey emotional content (see e.g., [Aryani et al., 2016]). This entire reasoning underpins the Gestalt approach to text presented here, viewing text as a cohesive entity comprising interconnected levels — sentences, words, and word parts — each playing a role in encoding emotional valence.

Preliminary studies and results

Investigating the idea of a relationship between the sublexical level of a text and its valency, a promising result was obtained of a correlation of individual vowel-consonant pairs with valency texts [Slavova 2019]. A corpus analysis was then

performed on the EmoBank corpus [EmoBank] containing 10,000 English sentences evaluated for valency, from 7 different genres (including newspaper headlines and travel brochures) [Buechel & Hahn, 2022]. The results derived from the corpus analysis [Slavova 2020, 2021] support surprisingly well the hypothesis that the emotional valence of text is also encoded in its phonemic composition. The system presented here employs the methodology applied in the corpus study, which is why it is briefly outlined here.

The sentences of the text decomposed into wordforms are presented at the sublexical level down to consonant-vowel pairs, in both forward and reverse order. These pairs are called *biphones*. The reasoning that led to the choice of such a sublexical "unit" is that in order to form words in speech, it is necessary to combine a vowel and a consonant.

To obtain the sublexical representation, biphone models of two types, consonant-vowel and vowel-consonant, were constructed, obtained as elements of a Cartesian product of the set of all vowels and that of all consonants as they are defined in English phonetics.

Phonetic transcriptions of words according to British standards are used. The transcribed words are broken down into bifones, and the sublexical level investigated further is that of bifones. The proposed methodology presents the sublexical level of the phrase "…*by another high school*...", transcribed as " *baɪ əˈnʌðə(r) haɪ skuːl* ", by breaking it down into constituent bifones such as: " *baɪ, ən, nʌ, ʌð, ðə, ər, haɪ, kuː, uːl* ".

The corpus analysis showed that the frequencies of many bifones in positive and negative sentences differ significantly. The principle component analysis of

the statistical space of bifone frequencies shows that the second principle component correlates with readers' valence ratings at 0.96, which is practically equivalent to emotional valence expressed through bifones [Slavova, 2021].

The results from corpus analysis confirmed that information related to emotional valence is encoded at the sublexical level. The discovered dependencies do not depend on the meaning of the utterance and are derived based on sets of unrelated sentences. The question of whether these dependencies apply to semantically connected text outside the corpus is crucial for clarification.

It was essential to validate the relationship between biphonic composition and emotional valence beyond EmoBank. Correlation coefficients (Pearson) between biphones and valence were calculated using textual material from the corpus and applied to texts outside the corpus. These coefficients range from -0.8 to +0.8, indicating association with positive or negative valence. However, finding a sufficient number of short, complete and properly evaluated texts for valence proved challenging, as open-access Internet resources lacked such material.

It was necessary to gather short and meaningful texts in English such as the news texts from prominent agencies like CNN, The Independent, etc. To evaluate them for valence, we use an alternative method. Using available dictionaries of emotionally rated English words, a list of the 100 most strongly negative (e.g., horrible, kidnapped, terror) and 100 most strongly positive (e.g., love, peace, happiness) words was compiled as keywords for headline searches. It was statistically verified based on experimental data from readers evaluation that the presence of negative or positive word in the headline

indicates very reliably the emotional tone the text. The presented here system uses these lists of highly emotional words to retrieve news items.

The task of gathering the news items from the Internet and their decomposition and analysis was performed by the system that we present here. Collecting and processing the data necessitates the development of a sophisticated system. Typically, such labor-intensive systems are implemented by multiple participants in the process. The system utilized in the latest study [Slavova & Andonov, 2022] was developed incrementally over several consecutive years, with involvement from undergraduate Computer Science students at New Bulgarian University (NBU). The curriculum includes the courses "Practice of Programming and Internet Technologies " and "Practice of Programming and Database Implementation", within which the students' participation was implemented. The methodology in the mentioned practices requires students, in groups of 3-5 people, to work in teams to implement a development according to an assignment from the teacher.

General presentation of the system

The purpose of the system is to gather data, which is later used to verify the hypothesis that there is correlation between the phonemic content of the text and its emotional valence. The system follows a seven-step process, in which texts are gathered from well distinguished news sources online, mainly American or English press agencies (The Guardian, ESPN, CNN, BBC, etc.).

As illustrated in Fig. 1, the system consists of two main components, a database and text processing modules. The system operates on exchanges between the two components. The input and output of each module is organized with storage in files, so that intermediate results can be passed to the next module to be written to the database, and the text processing process can be easily tracked.

Module A "collects" news texts by keyword in the headline. The texts are cleaned of links and stored, and their data is recorded in a DB table. Module B is a text parser, using a ready-made product for detecting names, numerals, dates, toponyms, etc. The module marks in the text the detected names, etc. with a special token, so that they are not taken into account when decomposing to biphones, filling as well a stop-word dictionary in the database.

Module C decomposes the texts to sentences and to wordforms, with identifiers and marked order respectively, recording the result both in a control file and in the database. A query retrieves the words that are not found transcribed in the database (the dictionary of transcribed word) and feeds them to the input of module D. Module D consults the Oxford online dictionary and records the transcriptions found in the database. Module E is the biphonic parser; it decomposes the newly found words into biphones and records them in the corresponding table in the database. Module F, query-based, generates the overall sublexical representation of the texts as biphones. Module G computes the weights of the decomposed texts as the sum of the weights of the participating biphones, normalized by the length of the text.

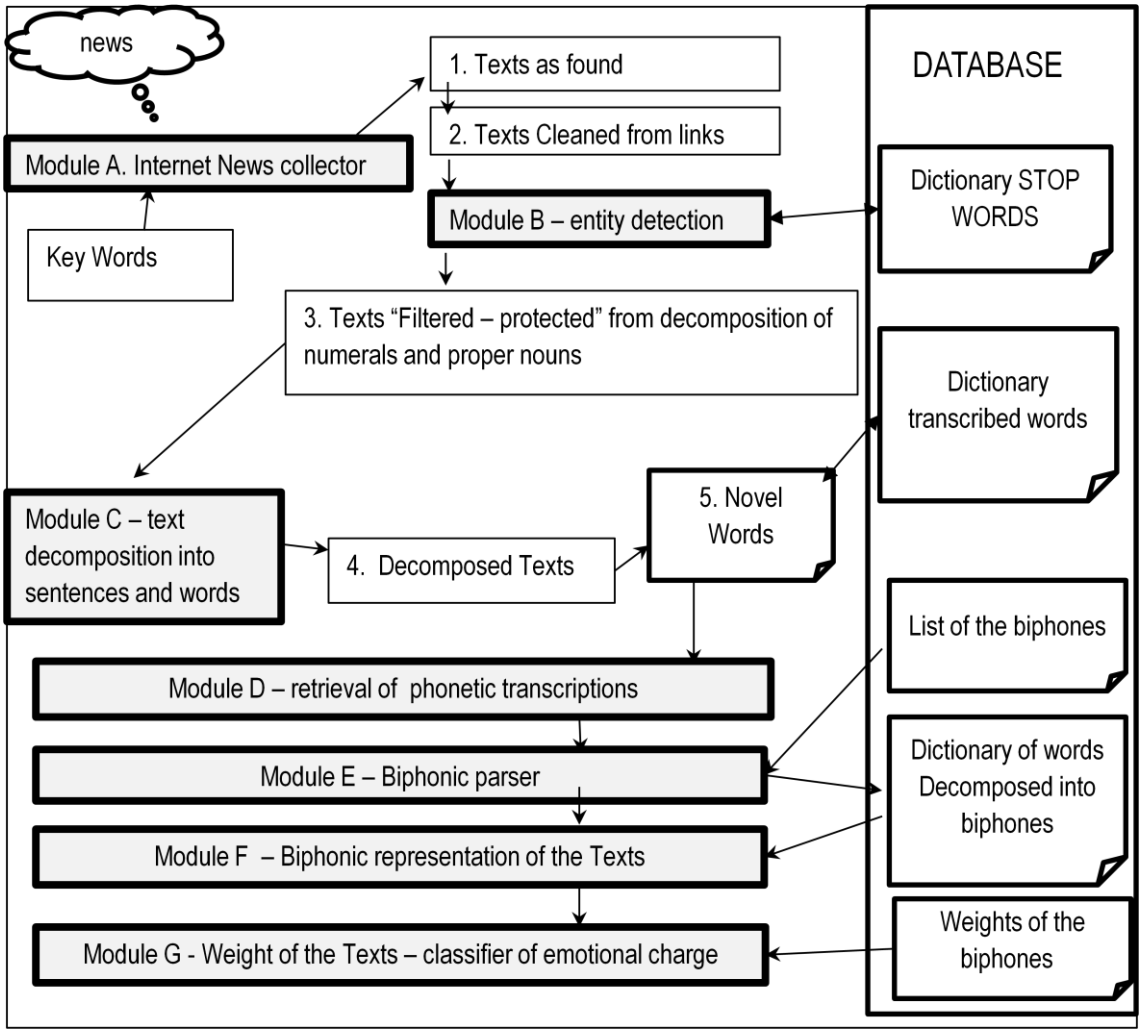

Figure 1. General structure of the modular system

Table 1 provides an illustration of a sentence from the downloaded news items, decomposed by the system to its biphonic composition.

The seven steps (the modules A – G) outline seven different tasks, each of which will be explained below in great detail. In general, each word in the gathered news texts is split into an array of biphones. The biphones are described by decomposing the wordforms' phonetic transcription, retrieved directly from Oxford Learner's Dictionaires [Oxford Dic]

Table 1: Example of a sentence from the news items, decomposed to biphones

Sentence	One of the seniors she's assigned to is 90 year - old Josephine Toia.				
word	position	Transcript	sequence	Biphone	BiphoneN
One	0	wʌn	1	wʌ	biph0965
One	0	wʌn	2	ʌn	biph0927
of	1	əv	1	əv	biph0413
the	2	'ðə	1	ðə	biph0208
seniors	3	ˈsiːniərs	1	siː	biph0751
seniors	3	ˈsiːniərs	2	iːn	biph0501
seniors	3	ˈsiːniərs	3	ni	biph0662
seniors	3	ˈsiːniərs	4	ər	biph0376
shes	4	ʃiːz	1	ʃiː	biph0335
shes	4	ʃiːz	2	iːz	biph0510
assigned	5	ə'saɪnd	1	əs	biph0378
assigned	5	ə'saɪnd	2	saɪ	biph0740
assigned	5	ə'saɪnd	3	aɪn	biph0069
to	6	tuː	1	tuː	biph0799
is	7	ɪz	1	ɪz	biph0988
90	8	Numeral			
year	9	jɪə(r)	1	jɪə	biph0576
year	9	jɪə(r)	2	ɪər	biph0533
old	10	əʊld	1	əʊl	biph0397
JosephineToia	11	Name			

From a technical point of view, the back-end of the platform is written entirely in Python, using the help of different libraries for web development, statistics and text processing. Here is a list of all the libraries used in the project:

XlsxWriter==1.3.7; nltk==3.6.5; openpyxl==3.0.5; openpyxl==3.0.5; pandas==1.2.0; python-docx==0.8.11; scipy==1.6.0; urllib3==1.26.7; xlrd==1.2.0; xlwt==1.3.0; spacy; tweepy; requests; newspaper3k; flask; vaderSentiment; validators; ilock; dask

The front-end is composed of plain Javascript, HTML and CSS (Bootstrap). The system is deployed on an AWS EC2 instance. The web system serves less than

1000 requests per week, so we decided that a small general purpose instance will provide just enough computing power for our use case, and precisely for this reason we selected the t2.micro instance.

Description of the modules

Module A (Fig. 2) is used for downloading articles. It starts by accessing the file 'KeyWords.xlsx' containing predefined keywords which are divided into positive and negative. After that the module starts searching and downloading articles containing the given keywords and that are between 1 and 2.5 pages long from The Guardian and CNN. The downloaded articles are saved divided into positive or negative folders in the 'Texts as found input' directory.

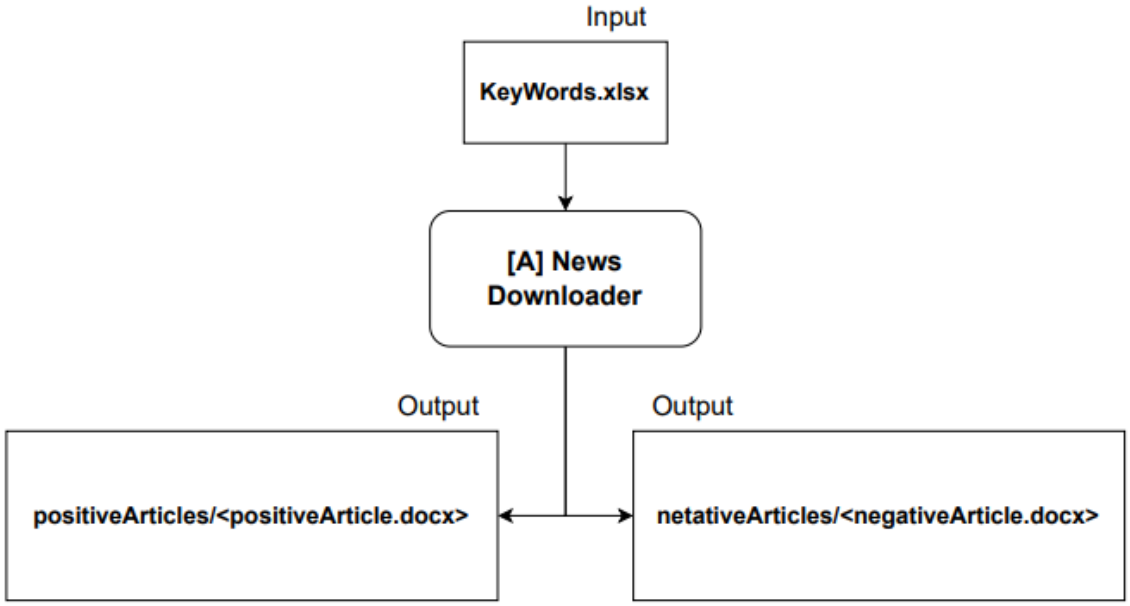

Figure 2. Module A

Module B is used for finding and marking all of the words which should not be further considered in the analysis such as names, numerals, dates, toponyms etc. They are marked in the text in order to indicate they should not be decomposed (see Table 1) and defined as Stop Words. It starts by iterating over all of the articles and using the spaCy's nlp, the module finds the stop words in the text, removes any whitespaces and unwanted characters and appends '~' at the end of every stop word. After that the file with the stop words marked is saved for later processing by the other modules in the 'TextsFiltered–protected' directory. The module also saves in an excel file all of the stop words of a given article in the 'Stop words' directory and, as shown in Fig. 3, at the end merges all of the separate excel files into one file containing all of the stop words of the given run.

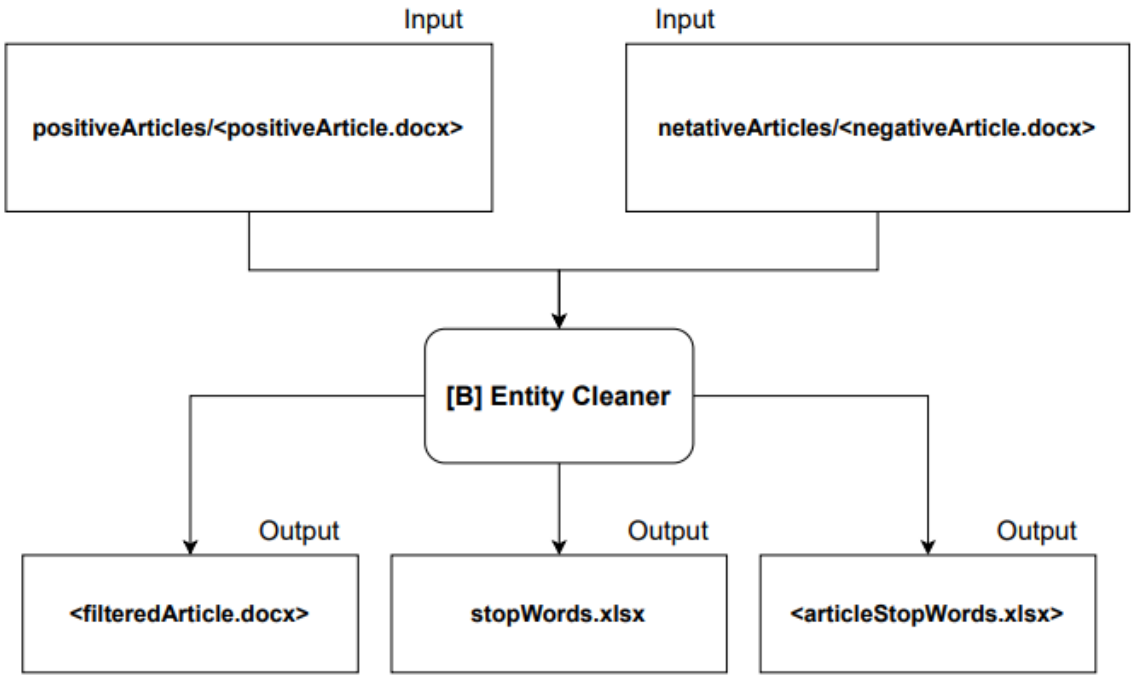

Figure 3. Module B

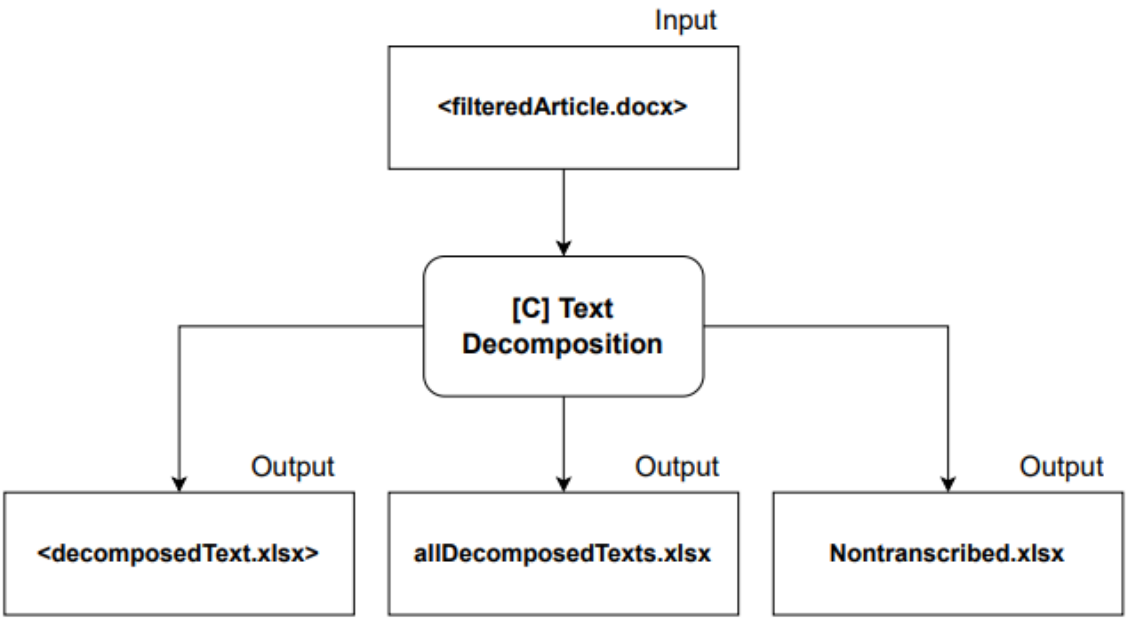

Figure 4. Module C

Module C, shown on Fig. 4, is used for decomposing the processed texts from module B into sentences and words. It starts by iterating over all of the texts and for every text file creates an excel file in the 'Decomposed Texts' directory in which the module decomposes the text into sentences (worksheet: Decomposition into sentences) and then in words (worksheet: Decomposition into words), After that all of the files with decomposed texts are merged together in a single excel file and adds all of the decomposed words in the database in the table 'words'. At the end the module makes a query that saves all of the decomposed words that are not stop words and do not already exist in the 'dictionary' table in the 'Non Transcribed.xlsx' file which is contained in the 'Novel Word' directory.

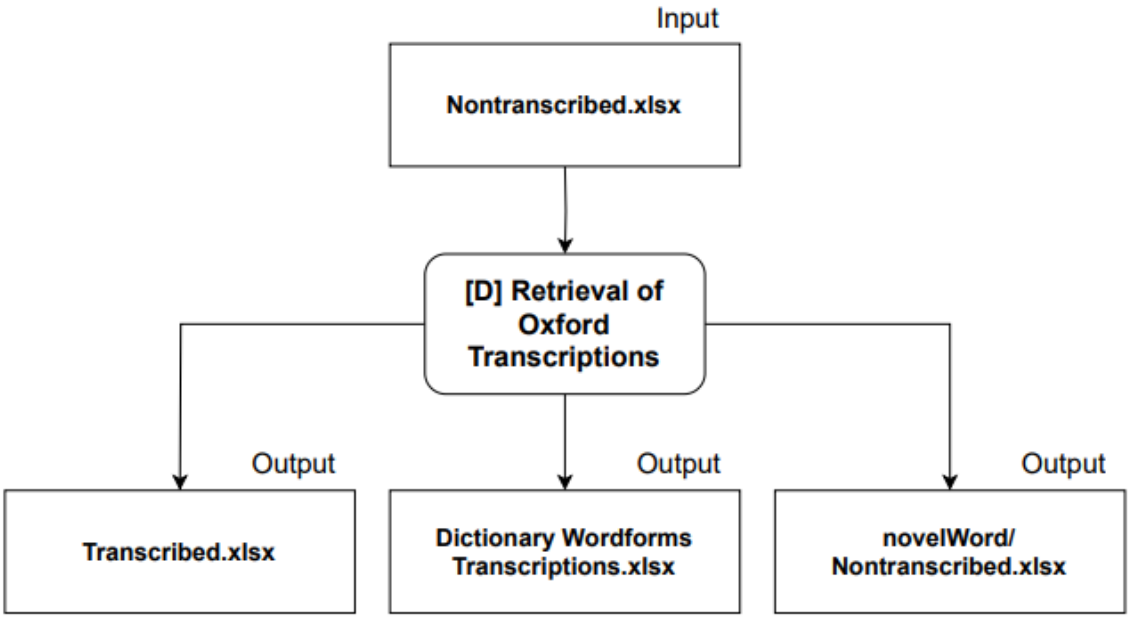

Figure 5. Module D

Module D (Fig. 5) is used for transcribing the non transcribed words that are saved in 'Novel Word/Non Transcribed.xlsx' by searching for these words in Oxford Learner's Dictionary. The module in itself defines some transcriptions that can't be found in the source. All of the words that the module successfully transcribes are saved in 'Novel Words/ Transcribed.xlsx' these new words are then added to the already existent data of transcribed words that is in the 'dictionary' table in the database and in the file 'DICTIONARY Wordforms Transciptions.xlsx' located in the 'Dictionary' directory. All of the words for which the module failed to find transcriptions are saved as stop words in 'Dictionary/ Stop Words.xlsx' and in the database in the table 'stop_words'.

Module E is used for decomposing the newly found words into biphones. The module begins by accessing the file 'Novel Words/ Transcribed.xlsx', as shown on Fig. 6. The newly transcribed words are cleared of any unwanted symbols like brackets, accents the space is replaced by underscore after which the module decomposes the words into biphones. The result of this decomposition is saved in 'Dictionary of words Decomposed Into Biphones/Words Contain Syllables.xlsx' and in the database in the table 'words_syllables'.

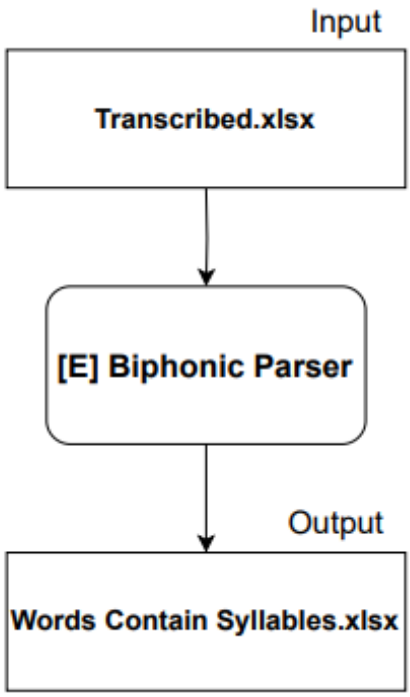

Figure 6. Module E

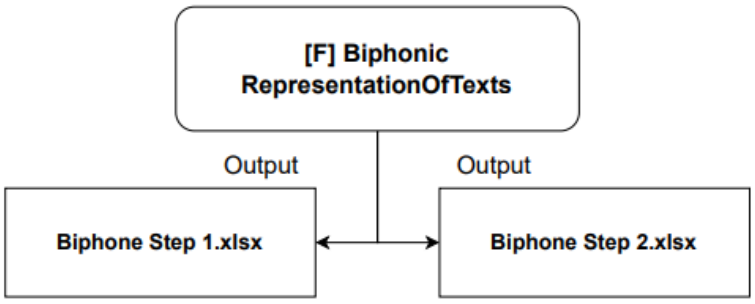

Figure 7. Module F

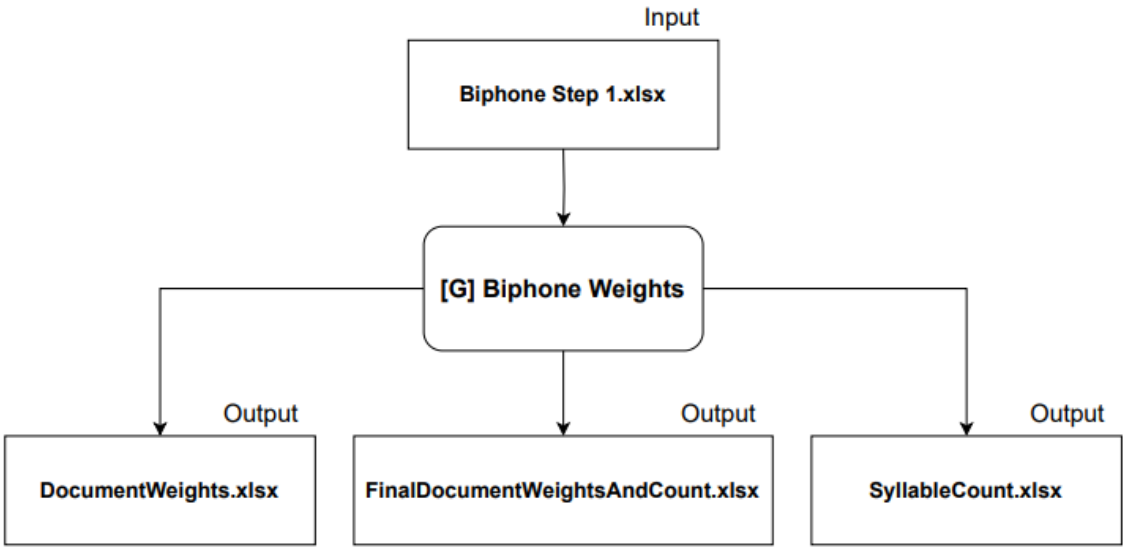

Figure 8. Module G

Module F executes two SQL queries that populate the 'Biphone Step 1.xlsx' and 'Biphone Step 2.xlsx' files with data. The data that is saved in 'Biphone Step 1.xlsx' is: each word from each file is separated on a new line providing information about the position it is in, what the word is, whether it is a stop word

or not. When the word is not a stop word, its transcription is given. The data recorded in 'Biphone Step 2.xlsx' is: each word from each file is separated into a new line or lines, the number of lines depends on how many syllables the given word is divided into, contains information about the sentence and the position in it of each word, the transcription of each word and information on each individual syllable.

The last module, Module G (Fig. 8), is used to compute the weights of the texts: To accomplish this it uses the data generated in 'Biphone Step 2.xlsx' and data for the biphones in the database (their correlation coefficients derived from EmobBank). It outputs three files: 'Document weights.xlsx', 'SyllableCount.xlsx' and 'Final document weight and count.xlsx'. Document weights contains information for each individual article: what is the r_value of the article, the number of words in it and the weight. SyllableCount contains information about the NonTransscribed words and NBiphones in each article, about the Transcribed words and about the result which is calculated by dividing the weight from the Document weights of the NBiphones. The Final document weight and count file contains the information from the previous two files combined.

As seen from the detailed description, the system is conceived to be easily tracked for inconsistencies. The system's intermediate results were rigorously checked step by step, and the inconsistencies – ruled out.

Conclusion

The developed system was used to statistically analyze the relationship between the sublexical representation of the phoneme composition and the

emotional tone of the news. The results obtained so far show that the phoneme composition-emotional valence relationships are manifested in about a page long meaning-complete texts [Slavova, Andonov, 2021, 2022].

The area of research is not close to computer science and the tasks, as well as the goal, are not intuitively understandable to non-specialists in natural language processing. Most of the students interested in the topic consulted recommended articles from specialized fields. The work required students to participate in extended meetings to monitor the outcome of performance, detect and discuss problems at all levels. Undergraduate students successfully complete complex work, find original solutions, and show interest in the results in the context of research.

AKNOWLEDGMENTS

The modules shown in Fig. 1 were developed with the active participation of student teams of third-year undergraduate students from several successive classes. We express our gratitude to all students participating in the teams that completed their work with working software modules and detailed documentation of their development. We cannot enumerate all the participants in the realization of the tasks. We would like to thank here the developers who participated very vigorously and efficiently in the last stages: Yvonne Tsonkova, Eric Baliov, Martin Bukreev, Mihail Zdravkov, Natali Arabadzhiyska, Katerina Koeva, Evelina Aleksandrova, Nikoleta Todorova.

Bibliography

[Aryani et al, 2016] Aryani, A., Kraxenberger, M., Ullrich, S., Jacobs, A. M., & Conrad, M. Measuring the basic affective tone of poems via phonological saliency and iconicity. Psychology of Aesthetics, Creativity, and the Arts, 10(2), 191. https://doi.org/10.1037/aca0000033

[Buechel and Hahn, 2022] Buechel, S., & Hahn, U. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. arXiv preprint arXiv:2205.01996.

[Kawahara & Shinohara, 2012] Kawahara, S., & Shinohara, K. A tripartite trans-modal relationship among sounds, shapes and emotions: A case of abrupt modulation. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 34, No. 34). https://escholarship.org/uc/item/47b452vw

[Slavova & Andonov, 2021] Slavova, V., & Andonov, F. How deeply are emotions encoded in language communication and is this detectable in text. International Journal "Information Theories and Applications, 1(3), 271-299. https://doi.org/10.54521/ijita28-03-p04

[Slavova & Andonov, 2022] Slavova, V., & Andonov, F. Bad news or good news when recognizing emotional valence using phonemic content. In 2022 21st International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE. https://doi.org/10.1109/INFOTEH53737.2022.9751339

[Slavova, 2019] Slavova, V. Towards emotion recognition in texts–a sound-symbolic experiment. International Journal of Cognitive Research in Science,

Engineering and Education (IJCRSEE), 7(2), 41-51 https://doi.org/10.5937/IJCRSEE1902041S

[Slavova, 2020] Slavova, V. Emotional valence coded in the phonemic content–Statistical evidence based on corpus analysis. Cybernetics and Information Technologies, 20(2), 3-21. https://doi.org/10.2478/cait-2020-0012

[Slavova, 2021] Slavova, V. On the revealing the emotional valence in communication by text. Procedia Computer Science, 192, 1514-1523.

IInternet Resources

[EmoBank] GitHub. https://github.com/JULIELab/EmoBank

[OxfordDic] https://www.oxfordlearnersdictionaries.com/).

Authors' Information

Rosen Stefanov *– student in computer science at New Bulgarian University; e-mail: rkstefanov1@gmail.com*

Major Fields of Scientific Research: natural language processing, AI

Martin Konstantinov *– student in computer science at New Bulgarian University; e-mail: martin.konstantinov0610@gmail.com*

Major Fields of Scientific Research: natural language processing, multicriteria optimization and decision analysis

Velina Slavova *– prof. in computer science at New Bulgarian University; e-mail: vslavova@nbu.bg*

Major Fields of Scientific Research: models and cognitive approach to AI