Identification of explicit lexicon for opinion mining in the media

Liauchuk Veranika

Abstract—The paper deals with the problems of opinion mining in respect to the media discourse. The peculiarities of analytical media genres as the sources for opinion extraction are considered, the opposition and distinction between opinionative and non-opinionative phrases (opinions and facts) are analyzed. The paper presents quantitative survey on opinions in media discourse, discusses opinion lexicon and presents a classification of opinionative phrases in the discourse of mass-media. The results can be used in the algorithms and programs of automatic identification and extraction of opinion.

Keywords **—** explicit means, media discourse, opinion lexicon, opinion mining, opinionative phrases

I. INTRODUCTION

Opinions fulfill all spheres of human activity and sway the perception of reality, behavior, choices and judgments upon a particular topic. Due to the role of the media as one of the sources to gain knowledge and, accordingly, the influence on a recipient, research of different media genres is becoming more important.

Studies of different points of view, comments and ratings in the media as well as facts and documental information create an opportunity to understand an object or a situation under consideration in a more profound way. Tracking of opinion trends as such and separate opinions in particular is helpful while investigating public opinion tendencies in the media-discourse. It is also noted that opinions (together with evidence and beliefs) record valuable data, and thus, become both a prerequisite and a result of perception [1].

II. HISTORICAL OVERVIEW OF THE ISSUE

A. Previous surveys

As far as opinions are considered a result of intellectual activity of a human-being, they became a subject of investigation in different fields (psychology, logics, philosophy, theory of communication, etc.). There is a range of works (Aleksandrova, Wittgenstein, Dmitrovskaya, Zaliznyak, Karagodin, Pavilens) dedicated to the opposition of knowledge and opinion (in this case a particular opinion is regarded as a proposition that requires verification or argumentation unlike propositions of knowledge), a number of surveys dealing with opinionated predicates (Apresyan, Arutunova, Zadvornaya, Ivanova), which provide foundations for further surveys.

B. Opinion mining and Sentiment analysis

The domain of subjectivity analysis is being elaborated nowadays. It is studied in two main areas – Opinion mining and Sentiment analyses. The first aims at extracting and processing

V. Liauchuk, (e-mail: missing).

views in a broad meaning, while the latter intends to determine attitudes. In a way the problems of both are similar, that is why at times these terms are used interchangeably.

The surveys of customer reviews, film reviews or ratings about products or services dominate in both areas (L. Lee, B. Pang, H. Tang, J. Wiebe, E. Riloff), thus, aiming at identification of emotional attitudes. There is also a number of works dedicated to differentiation of opinions and judgments which do not lack an emotional component (Dmitrovskya, Pazelskaya, etc.). However, as Arutunova suggests "there are no distinct boundaries that divide mental and emotional spheres, will and wishes, perceptions and judgements, knowledge and beliefs in the inner life of a person" [2], emphasizing by this the correlation between opinions and judgments.

III. THE POTENTIAL OF MEDIA-DISCOURSE FOR OPINION MINING

A. Mass-media discourse peculiarities

The media have a unique capability for creation of cognitive trends and their manipulation. Being affiliated in the process of production of meanings, images and metaphors, in the process of mass information creation (content), the media sway public perception and behavior, control intentions and the way they are pronounced and disseminated.

Marking suggestive influence of the media on a mass recipient, Zheltuhina stresses that "this role is unlikely to change in the nearest centuries as speed, quality and quantity of the information sent and received (newspapers, magazines, radio, television, the Internet) is modifying together with the technological progress. So, the level of the media influence is increasing due to the interpretation of events and transformation of the information transmitted". [3]

Werlich calls the types of modern texts "matrixes of elements that state texts fixed in linguistic experience of speakers" [4]. Moreover, Brinker notes that "texts represent a conventional example of complex linguistic actions" [5]. In this case a text can be regarded a complex unit with typical contextual (situational), communicative, functional and structural traits.

B. Analytical genres of the media

Analytical genres of the mass-media attract a particular interest for opinion extraction and study: analytical article, commentary, press or letter overviews, etc. The feature that unites them is orientation towards "interpretation of facts, phenomenon, elaboration of author point-of-view" and at times "a clash of opposite reader opinions in one material, dwelling upon it and revealing a personal judgment" [6]. It is logical to admit, the majority of linguistic means and tools are aimed at implementation of an affecting function "through argumentation, demonstration and reasoning" [7].

IV. SOURCES FOR THE SURVEY

The survey presented was carried out on the base of articles of informational and analytical genres from the British newspapers "The Guardian", "The Daily Telegraph", "The Independent", "The Daily Mirror" and Belarusian editions "Zvyazda", "Novy Chas" and "Culture". Through the method of contextual analysis more than 3000 opinionative phrases were selected from the mentioned texts.

V. OPINIONATIVE CONTEXTS

A. *Opinionative phrases*

Wittgenstein distinguished opinions as propositions that need verification against knowledge that does not need such verification. According to Apresyan, "the main difference between opinion and knowledge is in the relation to the verity of a proposition that contains the sense of a phrase" [8]. He also suggests a number of opinion-predicates and fact-predicates. Dmitrovskaya adds the criterion of probability that is implied by a speaker in a phrase [9]. In the surveys of Zaliznyak, opinion are characterized as phrases that have argumentation, comparison as a source of perception, they can have different levels of certitude, and finally, all unverifiable propositions should be considered opinions [10].

B. *Opinions vs Facts*

Naturally, not all the phrases in media texts contain opinionative markers. It is important to distinguish between opinionative and non-opinionative phrases [11].

In our research we understand "opinion" as a proposition, by which some person shares a particular point of view on some object or situation. The obligatory component in the structure of an opinionative phrase is subjectivity.

There are several options by which opinions are opposed to facts (as verifiable objects, situations that are unbiased empirical knowledge):

- Author interpretations, logical arguments and conclusions to which another author can object. For example, a phrase "Ernest Hemingway received a Nobel prize for the novel "The Old Man and the Sea" is a fact. However, a phrase "The novel "The Old Man and the Sea" by Ernest Hemingway received a Nobel prize because the author managed to show unique experience" is opinion as another author can declare that "[it was] due to the representation of the strength of a human soul" or "[…] thanks to a particular writing style".
- Emotional component in the structure of a phrase. Then, a phrase "The novel "The Old Man and the Sea" is the best work by Ernest Hemingway" is an opinion since one can oppose to it a similar in structure but a different in contents phrase – "The novel "For Whom the Bells Tolls" is the best work by Ernest Hemingway".
- Modality: "Ernest Hemingway should have dedicated the book to the fisherman Santiago".
- Statements about capabilities: "Ernest Hemingway could describe a life in a small fishing village in a thousand pages".
- Statements about the future: "The novel will always be popular", etc.

C. *Opinion phrase*

One of the problems for extracting opinions is to define a minimum unit (a phrase or context) that is enough to perceive a particular opinion. In our survey we define an opinion phrase as a fragment of a text that contains a personal point of view (opinion of an author or another sender of information on a topic discussed) that is enough to perceive the sense of this point.

Such units can vary profoundly in their size:

- a part of a sentence;
- a whole sentence;
- several sentences;
- a passage or even the whole text (especially if an opinion is announced in a heading).

VI. OPINIONS IN THE MEDIA: QUANTITATIVE PARAMETERS

A. *Explicit and implicit opinions*

When mining opinions, one deals not only with positive, negative or neutral grades, but rather with the extraction of information, to be more precise: what and how somebody thinks of something. Whatever the purpose is, opinions are results of intellectual activity of a person, of something that is created, developed and preserved in a human's consciousness. That is why the first step is to build opinion lexicon that contains markers of intellectual, logical, etc. activity of a person.

In is important to note that the contents of every phrase (both opinionative and non-opinionative) differs by the level of explicitness. Correspondingly, there are explicit (obvious) and implicit (concealed, implied) statements. An explicit phrase is a statement which sense can be extracted reasoning from a surface structure, meaning of words it consists of, without additional transformations [12]. It is also right that explicit statements "are represented directly in the lexical and syntactical structure of a phrase" [13]. Explicit markers (words and constructions) are instrumental for the extraction, analysis and further classification of opinions [14].

B. *Quantitative characteristics of opinions for informational and analytical genres of media discourse*

One part of our survey considered the saturation of opinionative phrases in informational and analytical genres of media discourse in general. To fulfill this task, we considered how one fact (an accident) was presented and discussed in both genres in the media both in English and Belarusian (attack on "Charlie Hebdo" offices on the 7th of January, 2015). And thus, only 16% of statements in the texts of analytical genres in English and 37% in Belarusian contain exceptionally facts (Table 1). For comparison, the proportion of such statements in the texts of informational genres is 49% and 60% accordingly. Moreover, while in informational genres authors express opinion explicitly rarely (9% of phrases in the English language and 8% in Belarusian) and prefer to share opinions by quoting other participants of communication, in the texts of analytical genres the proportion of the phrases with explicit means by author is much higher (48% in English and 47% in Belarusian).

TABLE I. STATISTICAL DATA ON OPINIONATIVE PHRASES IN THE TEXTS OF INFORMATIONAL AND ANALYTICAL GENRES OF THE MEDIA (REGARDING THE SAME FACT)

English		**Belarusian**		**Option**
Information	***Analytical***	***Information***	***Analytical***	
49%	16%	60%	37%	The volume of non-opinionative phrases
51%	84%	40%	63%	The volume of opinionative phrases
59%	9%	66%	8%	The percentage of quotations within opinionative phrases
26%	48%	25%	47%	The percentage of opinionative phrases with explicit markers
15%	43%	9%	45%	The percentage of opinionative phrases with implicit markers

C. Quantitative characteristics of opinions for the genres "Comment" and "Opinion"

The peculiarity of a media text is in its secondary nature in respect of a founding text (a text itself or a real-life situation). This feature is especially urgent for the media genres "Comment" and "Opinion" since their aim is to represent author's opinion on some topic. Accordingly, the main peculiarity of mentioned genres is high saturation with opinion phrases and explicit opinionative phrases in particular (Table 2). The rest of a text is a description of a situation discussed or facts.

The survey showed that the majority of opinionative phrases in the media genres "Comment" and "Opinion" are fulfilled with explicit of means of opinion expression. The tendency is similar in both languages. However, in English implicit opinionative phrases are also rather frequent (in comparison with explicit – 43% to 57% accordingly). This means that authors express opinion allegorical with the help of presupposition, irony and other means (for example, the author of an article running about the referendum in Scotland made a hint about a particular status of the region as compared to Wales or Northern Ireland in the following way: "Scottish independence: There's a kind of magic in our United Kingdom").

TABLE II. STATISTICAL DATA ON OPINIONATIVE PHRASES OF MEDIA GENRES "COMMENT" AND "OPINION"

Option	English	Belarusian
The percentage of opinionative phrases with explicit markers	57 %	82 %
The percentage of opinionative phrases with implicit markers	43 %	18 %
The percentage of quotations within opinionative phrases	6,5 %	48 %
The percentage of the opinionative phrases of an author (explicit and implicit)	93,5 %	52 %
The proportion of explicit and implicit opinionative phrases of an author	53,2 % to 46,8 %	66 % to 34 %
The percentage of the explicit opinionative phrases of an author	88,4 %	41,5%
The percentage of quotations within explicit opinionative phrases	11,6%	58,5%

As one could notice, the percentage of quotations (a direct means to represent opinion in the media discourse) in English is only 6,5% in the genres concerned, while in the Belarusian language it reaches 48% of all the opinionative contexts. In this respect the survey showed that with these quotations authors tend to provide a detailed opinion of another person (an expert, a participant of a problem situation, etc.) who finally becomes a "co-author" of a text. Taking into consideration this qualitative trait of Belarusian media discourse, the proportion of opinionative phrases of an author makes 85%, which proves the point.

VII. OPINION LEXICON

Building a comprehensive data base, opinion lexicon, is an important task for creation of an accurate algorithm (and a tracking program) in the framework of opinion mining. Identification of explicit means of opinion expression is easier in comparison with the implicit ones, though still it is efficient due to their frequency on the texts. Thus, it is possible to extract a considerable array of useful information from opinion lexicon.

Below one can find examples of such opinion lexicon identified in the texts of analytical genre of the media (concerning the mentioned above accident).

A. Words and structures of interpretation

Within explicit means of opinion representation by an author words and structures that help to interpret an event dominate (35% of phrases in English and 63% in Belarusian).

1) Units of logical relations, explanation of patterns (Belarusian variants are omitted).

a) verbs: *explain, attribute (to)*

b) prepositions: *by (doing smth)*

c) colloquial particles

d) conjunctional structures with the meaning of consequence: *if... (then), when… (then), interpretation*

For e.g., "We explained that the artists working at a magazine printed drawings that made two men angry".

2) Markers with the semantics of reason

a) verbs and nouns: *cause*, *reason* (for the reason; the reason is…)

b) conjunctions: *because*, *for/as/since* (in the meaning of "because"), *due to*

For instance, "The reason is that it is a technique of conflict, not a cause".

3) Descriptions and definitions

a) verbs with the meaning "to represent, to depict, to imply": *mean, seem, represent, describe, see (as), view (as), symbolize*

b) cliché-structures: *the thing is*

c) prepositions: *to plus an subject* (to somebody something is…)

For e.g., "On the cover of what is meant to be an anti-establishment magazine, this just symbolises – at best – egalitarian bigotry".

4) Aim

a) verb: *aim at*

b) prepositions: *for*

For e.g., "[…] the assassins' bullets were aimed squarely at free speech itself".

5) Result

a) verbs: *result*, *lead to*

b) nouns: *result*

c) adverb: *generally*

For instance, "As a result, all advice at the time was for America not to universalise its response to 9/11 […]".

B. Markers of opinion

1) A number of units that only mark a phrase as opinion, though do not add any new semantic senses to the context.

a) verbs: *suggest, show, create*

b) structures with the meaning of enumeration: *in one regard*

c) parenthesis: *it must be said, to be fair, by the way, among others*

As an example, "To be fair, that article was simply guilty of the Endtimes hubris that affects us all […]".

2) Personal characteristics of an author.

There is a range of linguistic means by which the attitude of an author towards the topic is identified. All such contexts contain a pronoun in the first person, singular (I, for me, etc.).

a) words and structures that mark emotions and mood of an author: *I am afraid/appaled/heartbrocken/glad/aware, I like/love/worry, etc*.

For e.g., "I'm glad to see development of detailed efforts in this country at government and local authority level to address extremism".

By this an author emphasizes important for him details, and a recipient can get some information about a personality of the author. As a rule a clause or a separate sentence combined with the marker by sense contain an opinion on the object blamed or supported by the author.

b) A range of opinion verbs and structures that mark opinion contexts: *suppose, believe, think, find, say, agree, regard, consider, my point is, etc.*

The context becomes more personal as a recipient get information about the opinion of an author directly.

For example, "I think the surviving Charlie Hebdo journalists really had no alternative but to show some image of Muhammad on this week's cover".

C. Theories, ideas, recommendations

Authors of the media texts strive not only to present some interpretations, but also suggest some new ideas or pieces of advice considering possible solutions to the problem discussed. Up to 29% of the phrases in English and 18% in the Belarusian language have in its structure explicit opinionative means that mark contexts containing some concepts and recommendations. For example, "We must not allow the assassin's veto".

The most frequent are:

1) Modals

a) modal verb *can* in the meaning "be able" ("It [terrorism] can kill people and damage property") and "it is needed, it is vital" ("freedom of speech can only be absolute").

b) modal verbs *should, must*

c) modal verbs *need, have to, ought to*

d) modal structures: *it's important (to do smth)*

2) Question-answer bodies

3) Nouns: *question, problem, idea, sense*

4) Predicates with positive or negative grades

5) Verbs in imperative mood

D. Expectations about future

22% of phrases in English and 11% in Belarusian contain words and structures by which authors make predictions and forecasts about the topic or it separate aspects.

1) Verbs in the forms of future tenses

2) Oblique mood

3) Some adverbs: *will, would, might, may, perhaps, probably*

For e.g., “There will now be cries from the security services and parliament for more powers and more surveillance”.

VIII. Classification of opinions in the media

Intentional and semantic analysis of the opinionative phrases in the texts of the media discourse shows that there are three significant types of opinions. Each type helps to extract a particular information.

A. Opinions-interpretations

These opinionative phrases contain interpretations, comments, descriptions and characteristics of the discussed topic. For instance, “If much can be done from a legal and contractual side without marriage, then marriage loses all credibility”. Accordingly, a recipient is provided with different approaches, explanations and logical interconnections.

B. Opinions-theories

By the phrases of this type authors share ideas, conclusions, possible solutions to the problem, personal notions and concepts. For example, “Love shouldn't be completely unconditional, but it also shouldn't be a gun to the throat” or “Give the poor creatures a break!” Thus a recipient gets some recommendations and ideas which can be used in real life.

C. Opinions-prognosis

Authors can not only analyze facts, situations and trends, but also build their own prognosis and make predictions concerning the probable outcomes, changes, development of the topic or a situation as such. To illustrate this, “England will pay a high price for such arrant selfishness”.

D. Comments upon classification

1) It should be noticed that sometimes it is impossible to divide a phrase into several separate opinionative phrases. That is why a number of contexts should be considered a mixed type (“But it’s not just the national newspapers which should be considered in this debate because the blow will land hardest on local newspapers and micro news sites”).

2) A range of phrases of different types may contain judgments in its structure, while judgment is not the purpose of a phrase but only a means to identify separate elements in the whole phrase.

3) The phrases with markers of opinion (that do not imply additional meanings) should be classified according to the type of an opinion expressed in a clause or a separate sentence combined with the marker by sense).

IX. Conclusions

The survey allows making several conclusions:

1) Modern mass media serve as reviews and chronicles of the current event, sources of knowledge about the world, source of analytic data, recommendations, prognosis and are able to serve as opinion polls to a certain extent. The media have a unique capability for creation of cognitive trends and are instrumental for manipulation of public opinion.

2) Both informational and analytical genres of the media are saturated with opinionative phrases, especially, with the ones having explicit means of opinion expression in their structure.

3) To extract data from the media it is important to build a particular opinion lexicon, containing words, structures and markers of opinion. Explicit means are divided into several groups: interpretational (logical relations and explanation, reasoning, definition, aim, result),

markers of opinion (neutral by semantics but provide information about point of view of an author), ideas and recommendations (modals, question-answer bodies, particular predicates, etc.) and expectations (verbs in future tenses, particular adverbs, etc.).

4) Opinionative phrases in the media are classified according to the type of information that can be extracted: interpretations, theories, prognoses. In some cases contexts should be considered a mixed type.

5) The results can be useful (as data base, principles and goals) for the algorithms and programs of automatic identification and extraction of opinion.

REFERENCES

[1] O.S. Aleksandrova, "Cognitive status of opinion: abstract of dissertation", Moscow, 2001, p.3 (in Russian).
[2] N.D. Arutunova, "Cosider" and "see" (to the problem of mixed propositional attitudes)", Moscow, 1989, pp. 9 – 15 (in Russian).
[3] M.R. Zheltuhina, "Political and mass-media discourses: influence – perception – interpretation", Moscow: MAKS Press, 2003, p.49 (in Russian).
[4] E. Werlich, "Problems of text typology", St.-P., 1984, p.8 (in Russian).
[5] E.V. Akulova, "Genres of the German routine communication", Saratov, 2001, pp. 66 – 67 (in Russian).
[6] L.A. Mutovkin, "Printed media. Newspaper genres", unpublished (in Russian).
[7] M.S. Kardumyan, "Linguistic peculiarities of analytical type of mass-media discourse", Stavropol, 2011, p. 5 (in Russian).
[8] U.D. Apresyan, "System senses "know" and "consider" in the Russian language", Moscow, 2001, p.7 (in Russian).
[9] M.A. Dmitrovskaya, "Knowledge and opinion: image of the world, image of a person. Logical analysis of a language", Indrik, 2003, p. 15 (in Russian).
[10] A.A. Zaliznyak, "Polysemi in a language and the means of its representation", Moscow, 2006, p. 476 (in Russian).
[11] M. Tsytsarau, "Survey on Mining Subjective Data on the Web", unpublished.
[12] A.N. Baranov, "Hidden (implicit) statement in linguistic expertise of a text", unpublished (in Russian).
[13] I.M. Kobozeva, "Linguistic semantics", Moscow, 2010, p. 214 (in Russian).
[14] B. Liu, "Sentiment Analysis and Opinion Mining", California: Morgan & Claypool Publishers, 2012, p. 12.

Co-funded by the Tempus Programme of the European Union

This publication is the result of the project implementation: TEMPUS CERES: Centers of Excellence for young RESearchers. Reg.no.544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES

CERES