DOI 10.36622/VSTU.2023.26.3.026
УДК 004.942:004.056

ФАКТОРЫ И СПОСОБЫ ВЛИЯНИЯ НА РАСПРОСТРАНЕНИЕ ФЕЙКОВ В СОЦИАЛЬНЫХ СЕТЯХ

А.А. Караханова, В.И. Белоножкин

Социальные сети и новостные агентства все чаще публикуют фальшивые (фейковые) новости для увеличения читательской аудитории или в рамках информационной войны, поэтому проблема обнаружения фейков в социальных сетях становится все более актуальной. В статье рассмотрено влияние вредоносных и интеллектуальных узлов на распространение ложной информации с применением модели простого заражения сети. Исследована динамика перехода узлов в различные состояния – восприимчивое, принимающее и иммунизированное.Предложено решение, которое можно использовать для снижения вероятности распространения фейков на примере обнаружения и фильтрации кликбейтов, эффективность которого была подтверждена экспериментально.

Ключевые слова: ложная информация, фейк, социальная сеть, кликбейт, моделирование.

Введение

В информационной сфере фейк представляет собой намеренно искаженную или фальсифицированную информацию, распространяемую с целью введения в заблуждение ее потребителей.

К признакам фейка относят:

– недостоверность источников;

– неправдоподобность (несоответствие действительности, известным фактам).;

– отсутствие подтверждающих фактических данных;

– наличие логических ошибок, противоречий;

– неизвестность (анонимность) авторов;

– использование гипертрофированно эмоциональной лексики.

В последние годы наиболее эффективным каналом распространения фейков стали социальные сети и мессенджеры – наиболее популярные среди интернет-пользователей средства коммуникации и получения информации.

Ряд исследований процессов распространения фейков [1] показал, что существенную роль здесь играют вредоносные узлы, присутствующие в социальных сетях (например, боты), но теоретической проработки механизмов их влияния на эффективность дезинформации на текущий момент явно недостаточно.

Анализ влияния вредоносных и интеллектуальных узлов на распространение фейков в сети

Рассмотрим модель сети размера N со средней степенью:

$$z = \sum_k kp(k), \qquad (1)$$

где $p(k)$ — распределение степеней.

Данная сеть включает в себя три типа узлов (рис. 1, а):

– обычные – способные (в зависимости от обстоятельств) как переслать фейк другим узлам, так и не делать этого;

– вредоносные (зараженные) – готовые целенаправленно распространять фейки;

– интеллектуальные (умные) – не участвующие в распространении фейков.

Узлы могут иметь три возможных состояния:

– восприимчивое – готовность к получению сообщений (S);

– принимающее – передача фейка другим узлам после его получения (A);

– иммунизированное – отказ распространять фейк после его получения (I).

Восприимчивый узел может изменить свое состояние – стать либо принимающим с вероятностью p, либо иммунизированным с вероятностью (1 – p), пока число его принимающих соседей больше 0. Если

восприимчивый узел переходит в принимающее или иммунизированное состояние, он не меняет его до конца процесса моделирования.

Представленная модель подходит для описания некоторых онлайн-поведений, таких как репост сообщений в социальных сетях.

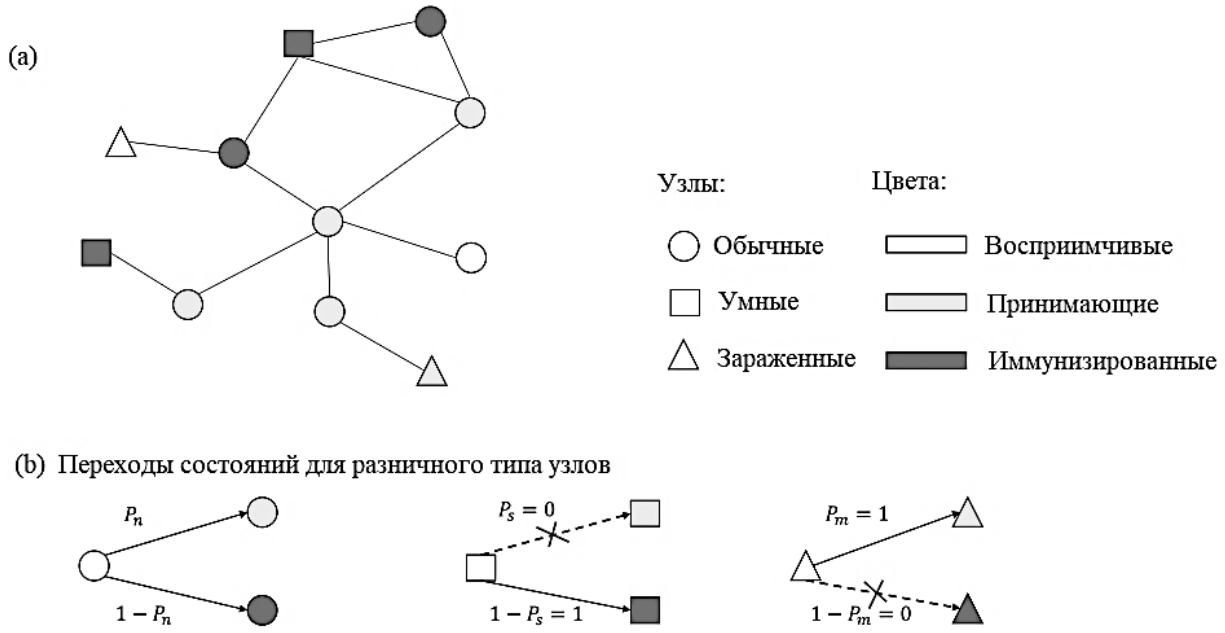

Рис. 1. Схематическое изображение исследуемой модели

Для исследования поведения модели случайным образом выбирается доля интеллектуальных узлов в сети (α), а также (случайно или следуя некоторым стратегиям, наиболее характерным для исследуемой сети) часть вредоносных узлов (β).

Разные типы узлов имеют разную вероятность перехода в состояние принятия при получении ложной информации. Пусть p_n, p_s и p_m будут вероятностями перехода в состояние принятия для обычных, интеллектуальных и вредоносных узлов соответственно. Предположим, что $p_s < p_n < p_m$.

Без потери общности примем, что $p_m = 1$, а $p_s = 0$, т. е. вредоносные узлы будут непременно ретранслировать заведомо ложную информацию, а интеллектуальные – не будут в любом случае.

Динамика процесса моделирования протекает следующим образом:

– первоначально предполагается, что все узлы находятся в состоянии S, за исключением одного узла (выбранного случайным образом) в состоянии A,

– на каждом временном шаге все узлы обновляются синхронно — узлы будут изменять свое состояние в соответствии с состояниями своих соседей на предыдущем временном шаге. В частности, восприимчивый узел изменит свое состояние, если количество принимающих соседей больше 0. Для разных типов узлов вероятности перехода из одного состояния в другое показаны на рис. 1(b). Узлы в состоянии A или I сохранят свои состояния неизменными,

– процесс завершается, когда больше не может быть изменено ни одного восприимчивого узла.

Рассмотрим случай, когда вредоносные узлы случайным образом распределены в сети с распределением степеней $p(k)$. и p_∞ – доля принимающих узлов в устойчивом состоянии.

Применяя анализ среднего поля, получаем:

$$p_\infty = p_0 + (1-p_0)\sum_{k=1}^{\infty} p(k)\left(\sum_{i=1}^{k} \alpha p_s b_{k,i}(p_\infty) + \sum_{i=1}^{k} \beta p_m b_{k,i}(p_\infty) + \sum_{i=1}^{k} (1-\alpha-\beta) p_n b_{k,i}(p_\infty)\right), \quad (2)$$

где $p_0 = 1/N$ – доля принимающих узлов в начальный момент времени,

$b_{k,i}(p_\infty) = \binom{k}{i} p_\infty^i (1-p_\infty)^{k-i}$ – это вероятность того, что узел степени k имеет i принимающих соседей в момент $t = \infty$.

Уравнение (2) можно понимать следующим образом: вероятность того, что случайно выбранный узел находится в состоянии A в стационарном режиме, суммируется из двух составляющих: вероятности того, что выбранный узел находится в состоянии A в момент $t = 0(p_0)$, и вероятности того, что узел восприимчив при $t = 0(1-p_0)$, но имеет по крайней мере одного принимающего соседа при t = ∞.

Учитывая все возможные типы выбранного узла, вероятность того, что он станет принимающим, равна

$$\sum_{i=1}^{k} \alpha p_s b_{k,i}(p_\infty) + \sum_{i=1}^{k} \beta p_m b_{k,i}(p_\infty) + \sum_{i=1}^{k} (1-\alpha-\beta) p_n b_{k,i}(p_\infty).$$

Сумма по k в уравнении (2) учитывает все возможные степени, которые может иметь узел.

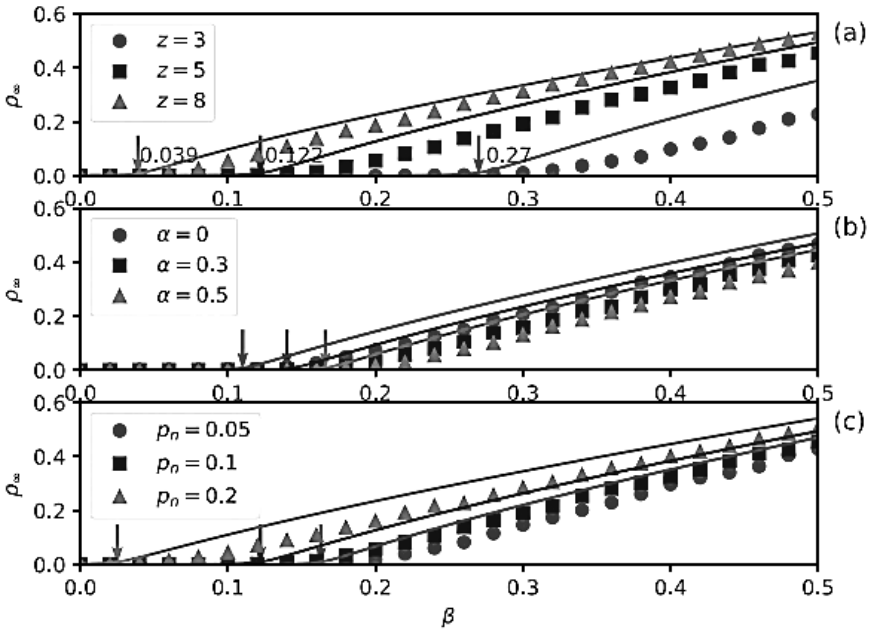

Рис. 2. Конечная доля принимающих узлов p_∞ как функция β для различных значений: a) средней степени z при параметрах N = 5000, α = 0.1 и p_n = 0.1; b) доля умных узлов α с параметрами N = 5000, z = 5 и p_n = 0,1; c) принятие вероятности для нормальных узлов p_n с параметрами N = 5000, z = 5 и α = 0.1. Стрелки указывают на теоретические значения β_c, предсказанные уравнением (5), а сплошные линии соответствуют теоретическим решениям уравнения (3) для разных значений β

Поскольку $\sum_{i=1}^{k} b_{k,i}(p_\infty) = 1 -$ упростим уравнение (2):

$$p_\infty = F(p_\infty) = p_0 + (1 - p_0) \sum_{k=1}^{\infty} [1 - (1 - p_\infty)^k] p(k) \times (\alpha p_s + \beta p_m + (1 - \alpha - \beta) p_n) . \quad (3)$$

Уравнение (3) можно решить численно для получения уровня распространения фейка (см. сплошные линии на рис. 2). Кроме того, из этого уравнения можно также определить условие возникновения ложной информации (имеется в виду, что $p_\infty > 0$ в пределе $N \to \infty$). Заметим, что $\rho_\infty = 0$ всегда является решением уравнения $p_\infty = F(p_\infty)$ (в случае $p_0 \to 0$, что верно при бесконечно большом N). Для получения положительного решения условие $F'(p_\infty)|p_\infty = 0 > 1$ должно выполняться, что приводит к:

$$\beta > \frac{\frac{1}{z} - (1 - \alpha) p_n - \alpha p_s}{p_m - p_n} . \quad (4)$$

Полагая $p_m = 1$ and $p_s = 0$, получаем:

$$\beta > \frac{\frac{1}{z} - (1 - \alpha) p_n}{1 - p_n} \equiv \beta_c . \quad (5)$$

Уравнение (5) показывает, что при заданных z, α и p_n существует пороговое значение доли вредоносных узлов β_c, выше которого ложная информация может распространяться по всей сети. Легко заметить, что β_c является убывающей функцией z, но возрастающей функцией α. Зависимость β_c от p_n представляется несколько более сложной. Вычисляя $\partial \beta_c / \partial p_n$ ($\beta_c + \alpha \leq 1$), можно заметить, что β_c будет уменьшаться при увеличении p_n.

Чтобы подтвердить теоретический анализ, было выполнено исследование выбранной модели в динамике. На рис. 2(a) показано, как конечная доля принимающих узлов p_∞ зависит от доли вредоносных узлов β для различных значений z. Продемонстрировано, что β_c уменьшается с увеличением z, что указывает на то, что распространение ложной информации становится легче. Теоретические значения β_c для различных значений z отмечены стрелками, как показано на рис. 2, a, что очень близко к результатам моделирования.

На рис. 2, b показана зависимость p_∞ от β для различных значений α – большему α соответствует большее β_c. То есть, чем больше в сети интеллектуальных узлов, тем потребуется больше вредоносных для успеха распространения фейков.

На рис. 2, c показано, что снижение доли обычных узлов, например, с помощью массового обучения пользователей, также может препятствовать распространению ложной информации в сети.

Способ выявления фейковых сообщений, использующих кликбейт

Одним из возможных направлений увеличения количества интеллектуальных узлов может быть распространение инструментов, позволяющих осуществлять обнаружение и фильтрацию потенциальных фейковых новостей.

Так, в число признаков фейка входит использование кликбейтов — фраз в заголовках сообщений, привлекающих внимание пользователей сенсационностью, эмоциональностью и побуждающих их перейти по по ссылке на требуемую веб-страницу.

Этот способ изначально использовался издателями веб-контента для увеличения доходов за счет роста количества «кликов» на рекламные страницы. Сейчас, как отмечают некоторые авторы, например, [2], кликбейт используется не только в маркетинговых целях, но и при создании фейков для привлечения пользователей к ложной и вредоносной информации [3].

Снизить негативные последствия применения кликбейта может использование инструмента, идентифицирующего и помечающего как недостоверные сайты из результатов поисковой выдачи или новостной ленты социальных сетей. Такой инструмент может загружаться

пользователем и добавляться в браузер или приложение для чтения новостных лент.

Инструмент может использовать различные методы, в том числе связанные с синтаксическими особенностями ссылок, чтобы определить, следует ли включать их в результаты поиска в процессе просмотра страниц, которые были извлечены поисковой системой. При этом должны выявляться страницы, ссылки на которые содержат фразы, вводящие в заблуждение (например, с большим количеством гипербол и сленговых выражений). Такие веб-страницы будут помечены как потенциальные источники ложной информации, а пользователь – уведомлен перед тем, как перейти на них.

Кликбейты, как правило, содержат значительно более длинные фразы, чем обычные сайты [4]. Для использования этого признака устанавливается порог длины фразы (экспериментально установленное значение – 8 слов). При превышении порога веб-страница помечается как возможный фейк.

Помимо синтаксических характеристик заголовков инструмент также должен отслеживать использование знаков препинания на веб-страницах, т. к. в заголовках фейков широко используются восклицательные и вопросительные знаки.

Поскольку кликбейты, как правило, ведут пользователей на веб-страницы, содержание которых сильно отличается от заголовка [5], большинство из них покидают эти страницы почти сразу после перехода, что формирует высокие показатели отказов, которые также можно использовать для идентификации фейков.

В контексте практического создания предложенного инструмента необходимо:

– сформировать базу данных кликбейтов на основе анализа контента соцсетей (например, ВКонтакте, Facebook, Reddit);

– вычислить атрибуты таких страниц и создать файлы данных для тренировки их обнаружения с помощью программного обеспечения для анализа данных и машинного обучения (например, свободный программный пакет WEKA) [6].

После сбора URL-адресов в базу данных был составлен скрипт на языке Python, который вычислял атрибуты из заголовка и содержимого веб-страниц. Особенности скрипта: поддержка мультиязычности; анализ заголовков, начинающихся с цифр; анализ контента, написанного заглавными буквами или содержащего вопросительные и восклицательные знаки; учет фактора ухода пользователей со страницы из-за несовпадения контента с заголовком.

Поскольку для WEKA требуется специально отформатированный ввод, в скрипте был использован алгоритм для извлечения необходимых параметров (рис. 3), Во всех экспериментах использовалась десятикратная перекрестная проверка.

```
1  Open URL file
2  for each title
3      title starts with number? 1 → outputfile
4      title contains ? and/or ! marks? 1 → outputfile
5      all words are capital in title? 1 → outputfile
6      users left the website after visiting? 1 → outputfile
7      contents have no words from title? 1 → outputfile
8      title contains keywords? NoKeywords → outputfile
9  end for
```

Рис. 3. Алгоритм вычисления атрибутов сайтов с фейковыми новостями

После считывания файла атрибутов страниц в WEKA было проведено их ранжирование на основе нескольких алгоритмов, чтобы выбрать наиболее релевантный для повышения точности и сокращения времени обучения:

1. Оценка атрибута информационного усиления – оценивает ценность атрибута,

измеряя прирост информации по отношению к классу.

$$InfoGain(Class, Attribute) = H(Class) - H(Class \vee Attribute). \quad (6)$$

По сути, этот алгоритм измеряет, как каждая функция способствует снижению общей энтропии H(X), которая определяется следующим образом:

$$H(X) = -\sum \left(Pi * log2(Pi)\right), \quad (7)$$

где *Pi* – вероятность класса *i* в наборе данных.

Энтропия в основном измеряет степень ”примеси” – чем ближе к 0, тем меньше примесей в наборе данных. Следовательно, хороший атрибут – это атрибут, который содержит больше всего информации, т. е. максимально уменьшает энтропию.

2. Оценка атрибута корреляции – оценивает ценность атрибута, измеряя корреляцию (Пирсона) между ним и классом. Номинальные атрибуты рассматриваются на основе каждого значения, которое учитывается как показатель. Общая корреляция для номинального атрибута достигается через средневзвешенное значение. Таким образом, индикатор значения номинального атрибута представляет собой числовой двоичный атрибут, который принимает значение 1, когда значение встречается в экземпляре, и 0 в противном случае.

В табл. 1 приведены результаты выбора атрибутов веб-страниц (которые были использованы в тестах) на основе атрибута «Прирост информации» и «Атрибут корреляции».

Таблица 1

Выбор атрибутов

Атрибут	Прирост информации	Атрибут корреляции
В начале цифра	0.0769	0.00432
В содержании есть слова из заголовка	0.776	0.00434
Содержит вопросительный и восклицательный знак	0.0863	0.00547
Все слова заглавными буквами	0.1196	0.105
Пользователь немедленно покинул веб-страницу	0.3673	0.12884
Ключевые слова	0.4456	0.27043

WEKA поставляется с классификаторами, которые используются для оценки данных и предоставления конечного результата. Среди них:

– классификатор байесовской сети (BayesNet): проводит обучение байесовской сети с использованием различных алгоритмов поиска и показателей качества, предоставляет структуры данных (структура сети, условные распределения вероятностей, а также средства, общие для алгоритмов обучения),

– логистический классификатор (Logistic): класс для построения и использования полиномиальной модели логистической регрессии с оценкой хребта,

– классификатор случайного дерева (Random Tree): класс для построения дерева, учитывающего K случайно выбранных атрибутов в каждом узле. Он не выполняет обрезку и имеет возможность разрешить оценку вероятностей классов (или целевого среднего в случае регрессии) на основе набора удержаний (обратная подгонка),

– классификатор наивного байесовского (Naive-Bayaes): класс для наивного байесовского классификатора, использующего классы-оценщики. Значения

точности числовой оценки выбираются на основе анализа обучающих данных. По этой причине классификатор не является обновляемым классификатором (который при обычном использовании инициализируется нулевыми обучающими экземплярами).

Для выбора наиболее эффективных классификаторов для конкретного набора данных были проведены эксперименты с последующей оценкой результатов классификации на основе показателей производительности (табл. 2).

– Точность – истинные срабатывания, деленные на предсказанные срабатывания (истинные срабатывания плюс ложные срабатывания).

– Отзыв – частота истинно положительных срабатываний, называемая также чувствительностью, которая представляет собой истинно положительные срабатывания, деленные на истинно положительные срабатывания плюс ложноотрицательные срабатывания.

– *F*-мера – комбинация точности и полноты. Для ее получения необходимо перемножить точность и полноту, затем разделить их на сумму точности и полноты и умножить на два.

– Площадь под *ROC*-кривой – часть координатной плоскости под графиком *ROC*-кривой, являющаяся мерой качества модели бинарной классификации в машинном обучении.

Логистический классификатор имеет самую высокую точность, 99,5%, и, следовательно, лучшее качество классификации.

Логистический классификатор и классификатор случайного дерева имеют наилучшую полноту (чувствительность) – 99,2%, а также превосходят другие по показателю *F*-меры. По показателю площади под *ROC*-кривой лучший результат дали байесовские классификаторы.

Таблица 2

Результаты классификации

Классификатор	Точность	Отзыв	F-мера	ROC
BayesNet	94.3%	97.1%	97.1%	100%
Logistic	99.5%	99.2%	99.4%	99.6%
RandomTree	99.2%	99.2%	99.4%	97.5%
Naive Bayes	98.6%	98.4%	98.5%	100%

Заключение

В рамках представленной работы был проведён анализ влияния различных типов узлов на распространение ложной информации в социальных сетях. Для этой цели была построена и исследована модель простого заражения сети, включающая вредоносные и интеллектуальные узлы.

Исследована динамика перехода узлов сети в различные состояния – восприимчивое, принимающее и иммунизированное с учетом различных факторов.

В работе также был предложен инструмент для снижения вероятности распространения фейков на примере обнаружения и фильтрации кликбейтов,

Инструмент может использовать различные методы, в том числе связанные с синтаксическими особенностями ссылок, чтобы определить, следует ли включать их в результаты поиска в процессе просмотра страниц, которые были извлечены поисковой системой. Он позволяет также вычислять атрибуты страниц и создать файлы данных для тренировки их обнаружения с помощью программного обеспечения для анализа данных и машинного обучения.

Эффективность предложенного инструмента была подтверждена экспериментально.

Список литературы

1. How to combat fake news and disinformation / URL: https://www.brooking-s.edu/research/how-to-combat-fake-news-and-disinformation/ (дата обращения: 24.08.2023).

2. A simple but tough-to-beat baseline for the fake news challenge stance detection task. /

URL: https://arxiv.org/abs/1707.03264 (дата обращения 24.08.2023)

3. Lewis, S., 2011. Journalists, social media, and the use of humor on twitter. The Electronic Journal of Communication / La Revue Electronic de Communication 21, C. 1–2.

4. User Behavioral Analytics: The New Cybersecurity Approach / URL: https://www.-cybraics.com/blog/user-behavioral-analytics-the-new-cybersecurity-approach: (Дата обращения 24.08.2023)

5. Spicer, R.N., 2018. Lies, Damn Lies, Alternative Facts, Fake News, Propaganda, Pinocchios, Pants on Fire, Disinformation, Misinformation, Post-Truth, Data, and Statistics. Springer International Publishing, Cham. P. 1–31.

6. Free software // Национальная библиотека им. Н. Э. Баумана Bauman National Library / URL: https://ru.bmstu.wiki/Free_software (дата обращения: 24.08.2023).

Воронежский государственный технический университет
Voronezh state technical university

Поступила в редакцию 5.09.2023

Информация об авторах

Караханова Анна Александровна – аспирант, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Белоножкин Владимир Иванович – д-р техн. наук, профессор, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

MEANS AND METHODS OF MONITORING AND PREVENTION OF THE DISTRIBUTION OF FALSE INFORMATION

A.A. Karakhanova, V.I. Belonozhkin

Social networks and news agencies are increasingly publishing fake news to increase readership or as part of psychological warfare, so the problem of detecting fakes in social networks is only gaining relevance. The article considers the impact of malicious nodes on the spread of false information using a simple infection model with the inclusion of intelligent nodes that recognize false information better than ordinary nodes. The dynamics of the transition of nodes to various states - susceptible, receiving and immunized - was investigated. A solution has been proposed that can be used to reduce the likelihood of the spread of fakes using the example of detecting and filtering clickbates, the effectiveness of which has been confirmed experimentally.

Keywords: false information, fake, social network, clickbait,

Submitted 5.09.2023

Information about authors

Anna A. Karakhanova – Graduate Student, Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com

Vladimir I. Belonozhkin – Dr. Sc. (Technical), Professor of Voronezh State Technical University, e-mail: alexanderostapenkoias@gmail.com